The Dissertation Committee for David William Palmer Certifies that this is the approved version of the following dissertation:

Departing From Frankfurt:
Moral Responsibility and Alternative Possibilities

Committee:

Robert Kane, Co-Supervisor

John Deigh, Co-Supervisor

John Martin Fischer

Carl Ginet

Stephen White

Paul Woodruff

Departing From Frankfurt:
Moral Responsibility and Alternative Possibilities

by

David William Palmer, B.Sc.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2009

Acknowledgements

Many people have helped me with my dissertation. All of the members of my dissertation committee provided valuable input at various stages of the project. I should particularly like to mention the help of John Deigh, Carl Ginet, and Robert Kane. Professor Deigh worked through many drafts of chapters with me, patiently asking questions and suggesting good revisions. Professor Ginet read my work with far more detail that could have been expected for an outside reader. I consider him a *de facto* third co-supervisor. I am particularly grateful to him for this, and for inviting me to work with him on a jointly authored paper. Finally, Professor Kane has been my mentor from the beginning. All that I have learned about how to do good philosophy is directly traceable to him. He is an inspiration to me. Finally, I should also like to thank collectively my friends, family, and loved ones, for their patience and encouragement. I appreciate it enormously.

Departing From Frankfurt:
Moral Responsibility and Alternative Possibilities

Publication No._____

David William Palmer, Ph.D.

The University of Texas at Austin, 2009

Supervisors: Robert Kane and John Deigh

One of the most significant questions in ethics is this: under what conditions are people morally responsible for what they do? Assuming that people can only be praised or blamed for actions they perform of their own free will, the particular question that interests me is how we should understand the nature of this freedom – with what *kind* of freedom must people act, if they are to be morally responsible for what they do?

A natural answer to this question – and the one I think is correct – is to point to *the freedom to do otherwise*. This is encapsulated in the *principle of alternative possibilities (PAP),* the principle that a person is morally responsible for what he has done only if he could have done otherwise. PAP has led many to believe that the freedom required for moral responsibility must be incompatible with determinism or the existence of God because it is plausible to argue that if determinism is true or if God exists, then people would lack genuine freedom of choice and hence could not be morally responsible for their behavior.

In the light of two important articles by Harry Frankfurt almost four decades ago, which challenged the claim that moral responsibility requires the freedom to do otherwise, compatibilism – the opposing view that the freedom for moral responsibility *is* compatible with determinism – has experienced a resurgence. Inspired by Frankfurt's work, those wanting to reject PAP – typically compatibilists – attack the principle on two main grounds: directly and indirectly. First, they have argued *directly* that PAP is false by developing alleged counterexamples to it. Second, they have challenged PAP *indirectly* by arguing that there are alternative conceptions of freedom from freedom of choice that, it is claimed, are *not* reliant on alternative possibilities but *are* sufficient to capture the freedom required for moral responsibility.

My dissertation evaluates these two lines of attack on PAP. In particular, I attempt to defend the truth of PAP against both kinds of challenge.

Table of Contents

Chapter One: Moral Responsibility and Alternative Possibilities

1.1 The Central question of the dissertation.

Among the most significant questions in ethics is this: under what conditions are people morally responsible for what they do? Assuming that people can only be praised or blamed for actions they perform of their own free will, the particular question that interests me is how to understand the nature of this freedom – with what *kind* of freedom must people act if they are to be morally responsible for what they do?

A natural answer to this question – and the one I think correct – is to point to what we might call the *freedom of choice* or the *freedom to do otherwise*. This is the freedom to choose to perform one action or another, some action or none at all. It is encapsulated in the following principle:

> *The Principle of Alternative Possibilities (PAP):* A person is morally responsible for what he has done only if he could have done otherwise.

PAP has led many to believe that the freedom required for moral responsibility must be incompatible with causal determinism or with God's existence. For it is plausible to argue that if determinism is true or if God exists, then people would lack the freedom to do otherwise and so could not be morally responsible for their behavior.

Almost four decades ago, Harry Frankfurt published two papers, 'Alternate possibilities and moral responsibility' (1969) and 'Freedom of the will and the concept of a person' (1971), that dramatically changed the subsequent study of free will and moral responsibility. In the light of these two important articles, compatiblism – the opposing view that the freedom required for moral responsibility *is* compatible with determinism or

God's existence – has experienced a resurgence. Inspired by Frankfurt's work, those wanting to attack PAP – typically compatibilists – have done so on two main grounds: directly and indirectly.[1]

First, they have argued *directly* that PAP is false by developing alleged counterexamples to it, the most common of which are known as 'Frankfurt cases' (from Frankfurt, 1969). In these examples, a person acts apparently freely and responsibly despite being prevented from doing otherwise by a counterfactual intervener – that is, by someone who *would* have intervened *had* the person been about to act differently. Second, PAP rejecters have challenged the principle *indirectly* by arguing that there are alternative conceptions of freedom from the freedom to do otherwise that, they claim, are *not* reliant on alternative possibilities but *are* sufficient to capture the freedom required for moral responsibility. My dissertation, in the broadest strokes, is a defense of the claim that moral responsibility requires the freedom to do otherwise against these two lines of attack.

1.2 The Importance of the debate about PAP and Frankfurt's arguments.

Incompatibilists have tended to argue that moral responsibility is incompatible with the truth of determinism as follows:

---

1 Not all those wanting to reject PAP are compatibilists. Some incompatibilists, like Derk Pereboom (2001), argue that PAP is false because they ground their incompatibilism not on the claim that moral responsibility requires the freedom to do otherwise, but on the idea that it requires that the person is a genuine source of his behavior, something they believe the truth of determinism would eliminate. Also, not all modern compatibilists want to reject PAP. Some still follow the traditional compatibilist idea of interpreting the freedom to do otherwise in compatibilist-friendly ways.

The Traditional Incompatibilist Argument:

(1) A person is morally responsible for what he has done only if he could have done otherwise.

(2) The freedom to do otherwise is incompatible with the truth of causal determinism.

(3) Therefore, moral responsibility is incompatible with the truth of causal determinism.[2]

Up until Frankfurt's papers, compatibilists and incompatibilists largely agreed that premise (1) was true. That is, the truth of PAP was generally common ground. What was open to question, though, was premise (2), whether or not the freedom to do otherwise is incompatible with causal determinism. This premise has struck many as obvious and intuitive. More recently, incompatibilists argued for its truth by way of what has come to be called the *consequence argument* (see, for instance, Ginet, 1990 and van Inwagen, 1983 for formalized versions). Roughly, the argument is this. If determinism is true, then facts about the past and the laws of nature jointly entail facts about people's present actions. But since (i) people have no choice about either the facts about the past or the laws of nature, and (ii) if they have no choice about these, then they have no choice about things that are entailed by these facts, then (iii) if determinism is true, people have no choice – no freedom to do otherwise – with respect to their present actions.

However, the bulk of PAP-rejecters have been compatibilists and they will be my main focus in the dissertation.

2 A similar argument can be made to show that moral responsibility is incompatible with God's existence on the grounds that the freedom to do otherwise is incompatible with either God determining our behavior or His having complete foreknowledge of it.

Compatibilists, who prior to Frankfurt's papers had generally agreed with incompatibilists that PAP is true, typically responded to the traditional incompatibilist argument by claiming that the freedom to do otherwise, when properly understood, *is* compatible with the truth of determinism after all. To make this case, these so called *classical compatibilists* offered an analysis of the freedom to do otherwise in conditional terms (see Ayer, 1954 and Moore, 1912 for defenses of this view). According to the conditional analysis in its most promising form, to say that a person is free to do otherwise is to say that he *would* have done otherwise, *if* he had wanted to. Armed with this kind of analysis, compatibilists argued that the freedom to do otherwise *can* be properly analyzed in ways that do not conflict with determinism, and efforts to argue otherwise – by way of the consequence argument, for instance – fail once we understand the idea of 'having a choice' about something in compatibilist-friendly terms.

Such classical compatibilists then argued that premise (2) of the traditional incompatibilist argument is false because a person would be free to do otherwise in the conditional sense even if determinism were true. With respect to the incompatibilists' argument that premise (2) is true – the consequence argument – compatibilists argued that if we understand a person's 'having a choice' about something in these conditional terms, then the consequence argument is unsound. Even though people would not have a choice, in this sense, about facts about the past and the laws of nature (it is not true that if a person wanted to change facts about the past or the laws of nature, he would have done so), this would *not* entail that they do not have a choice, in the conditional sense, about their present actions. After all, it seems plausible to think that people have the freedom to do otherwise in the sense that they would have done otherwise, if they had wanted to.

The problem with these classical compatibilist responses is that the conditional analysis of the freedom to do otherwise on which they rest is untenable, as almost all now agree. Even many prominent contemporary compatibilists concede this point.[3] First, counterexamples have been offered designed to show that the conditional analysis is not *sufficient* for the freedom to do otherwise. Michael McKenna (2004) – following Roderick Chisholm (1964) – describes a case in which a young woman, Danielle, is psychologically incapable of wanting to touch a blond haired dog because of an earlier childhood trauma. He imagines that on her sixteenth birthday, her father presents her with a black haired dog and a blond haired dog and tells her that she must pick one to keep (she has been wanting a pet for some time). Danielle happily picks the black haired dog.

Was Danielle free to do otherwise and pick the blond haired dog? It seems not since her childhood trauma has left her psychologically incapable of forming a desire for blond haired dogs. So she was *not* free to do otherwise and pick that dog. Yet Danielle would appear to satisfy the conditions of the conditional analysis. It seems right to say of her that *if* she had wanted to pick the blond haired dog, she *would* have done so. Her problem, of course, is that, given the scarring effects of her childhood trauma, it is not psychologically *open* to her to want to pick that dog. So, satisfying the conditional analysis, as Danielle does, is not *sufficient* for having the freedom to do otherwise.[4]

---

3 Michael McKenna (2004) and John Martin Fischer (Fischer, Kane, Pereboom & Vargas, 2007), for instance, makes this concession.

4 One might try to avoid this problem by adding to the conditional analysis the requirement that it must be true to say of the person that she *could* have wanted to act differently (that, in Danielle's case, she *could* have wanted to pick the blond haired dog).

Second, J. L. Austin (1979) pointed out that what is often meant in saying that a person is free to do otherwise does not seem to be captured by saying that if he had desired to act in some other way, he would have acted in that way. By using the example of a golfer who tries but misses a putt, Austin argues that what we mean when we say that the golfer *could* have holed the putt (he was free to do otherwise than he did) is not that he would have holed it if conditions have been different (though that may be so). What we mean is that he could have holed it given the conditions as they were at that moment.[5] According to Austin, then, what we mean when we speak of someone as being free to do otherwise is not captured by the traditional compatibilist conditional analysis.

Some (e.g., Ekstrom, 2000 and Kane, 1996) argue that Austin's example shows that satisfying the conditional analysis is not *necessary* for having the freedom to do otherwise. Suppose in saying that the golfer could have done otherwise, we mean to imply that he could have holed the putt. According to the conditional analysis, to say he was free to hole the putt is to say that he would have holed it, if he had wanted to. However, this does not seem true of the golfer. As evidence, consider the fact that, in the actual circumstances, he *did* want to hole the putt but he did not do so – he missed. So it

---

But this just seems to push back the question of how to understand the term 'could have' to a higher level, i.e., what does it mean to say that a person 'could have wanted to act differently'? Alternatively, one might require that the person *not* be subject to psychological compulsion, hypnosis, subliminal advertising, and so on – factors which would prevent a person from being *able* to form a desire to act differently. But, as Fischer (Fischer, Kane, Pereboom & Vargas 2007: 51-52) – a prominent modern compatibilist – points out, there seems little chance of supplying a principled criterion for deciding what circumstances should go into such a list.

5 Austin (1979) writes, "Consider the case where I miss a very short putt and kick myself because I could have holed it. It is not that I should have holed it if I had tried: I did try, and missed. It is not that I should have holed it if conditions had been different: that

does not seem right to say that if he had desired to hole the putt, he would have done so. The golfer has the freedom to do otherwise and hole the putt (it seems) without satisfying the conditional analysis. This demonstrates that satisfying the conditional analysis is not necessary for people to be free to do otherwise.

The way this dialectic has unfolded has led many to believe that incompatibilism is in the stronger position than compatibilism. So long as all parties generally agree that moral responsibility requires the freedom to do otherwise, it has seemed to many more plausible to argue that this kind of freedom is incompatible with the truth of determinism or God's existence than to argue that it is not. Frankfurt's arguments against PAP are important, and have attracted so much attention, because they offer a new way for compatibilists to respond to the traditional incompatibilist argument. Armed with Frankfurt's arguments, compatibilists can simply deny premise (1), the claim that moral responsibility requires the freedom to do otherwise. Compatibilists would no longer be saddled with the burden of providing a convincing analysis of the freedom to do otherwise that is consistent with the truth of determinism. To this extent, compatibilism becomes a much more attractive option, should Frankfurt's attack on PAP be true.

If Frankfurt's argument against PAP is sound, then incompatibilists would be deprived of their traditional argument. It would not matter whether or not the freedom to do otherwise is incompatible with determinism, for this kind of freedom would not be required for moral responsibility. Incompatibilism would no longer occupy the intuitive

---

might be of course be so, but I am talking about conditions precisely as they were, asserting that I could have holed it. There is the rub" (p. 218, footnote 1).

'high ground' over compatibilism. To this extent, incompatibilism would become a much less appealing option.

Some incompatibilists suggest that not all would be lost in such a situation. Some (e.g., Pereboom, 2001) have argued that the success of Frankfurt's *direct* attack on PAP in particular need not undermine incompatibilism *per se*. These philosophers argue that the traditional incompatibilist argument should not be the main argument for incompatibilism. Moral responsibility does not require the freedom to do otherwise, they insist, but rather the freedom to be the genuine source of one's actions. Consider the following argument:

The Source Incompatibilist Argument:

(1) A person is morally responsible for what he has done only if he is the genuine source of his actions.

(2) The freedom to be a genuine source of one's actions is incompatible with the truth of causal determinism.

(3) Therefore, moral responsibility is incompatible with the truth of causal determinism.

So called 'source' incompatibilists argue that Frankfurt's direct argument, if successful, would not undermine incompatibilism properly construed. In fact, these incompatibilists may well be amenable to his direct attack on PAP to the extent that it would help shift incompatibilism *away* from the thought that moral responsibility requires the freedom to opt between alternative courses of action and *to* the claim they prioritize, the claim that a person is morally responsible for what he has done only if he is the genuine source of his behavior. Of course, these incompatibilists would not be in

complete agreement with Frankfurt's *indirect* attack on PAP. They would agree with him that there is an alternative conception of freedom from freedom of choice that is sufficient to capture the freedom required for moral responsibility. However, they part company from Frankfurt in believing that this freedom is not compatible with the truth of determinism.[6]

Having motivated the importance of the debate surrounding PAP, I now outline my defense of the principle as it proceeds in each chapter of my dissertation.

1.3 Summary of chapters.

Chapter two – 'Frankfurt's Direct Attack on PAP – The Frankfurt Cases.'

In chapter two, I introduce the Frankfurt cases. Consider the following example, similar to one Frankfurt develops in 'Alternate possibilities and moral responsibility' (1969). Imagine a person, Jones, decides to lie to save embarrassing himself, despite knowing that it is morally wrong for him to do this. Unbeknownst to Jones, he is unable to do otherwise because of the presence of Black. Black is a counterfactual intervener, someone who could and would intervene to make Jones lie, if Jones had not decided to lie 'on his own.' But, since Jones acted 'on his own,' without Black needing to intervene, it seems right to think that Jones is blameworthy for lying despite lacking the freedom to do otherwise. Hence, PAP is false.

---

6 I am inclined to reject source incompatibilism and think that incompatibilism should be argued for via the truth of PAP rather than appealing to the idea of a person being the genuine source of his actions. I offer no direct argument for this claim in the dissertation. However, I argue for it, albeit obliquely and incompletely, in chapters six and seven by suggesting that we can make good sense of the idea that a person is a genuine source, or author, of his behavior even if determinism were true.

I reject such cases because I think they fall prey to a dilemma, both horns of which undermine their cogency. Either Jones' decision to lie is causally determined or it is not. On the one hand, if it *is* determined, then the example begs the question against the incompatibilist who does not think that determinism is compatible with moral responsibility from the outset. On the other hand, if Jones' decision to lie is *not* causally determined, then Jones would have alternative possibilities after all, as PAP requires. This is because, since Jones' decision is not determined, Black will have to wait until Jones has begun to decide one way or the other, to lie or not to lie, to know whether or not he needs to intervene. But this waiting allows Jones to have alternative possibilities at the moment of choice, lying or not lying. And if Black should intervene *before* Jones has begun to decide, then Black, but not Jones, would be morally responsible for the decision.

This dilemma was originally developed by Robert Kane (1985), Carl Ginet (1996), and David Widerker (1995). Since then, those wanting to reject PAP have responded in two ways. Some (e.g., Fischer 1999, 2006 and Haji & McKenna 2004) have argued that a causally determined Frankfurt case *can* be used to show that PAP is false without begging the question against the incompatibilist. Others (e.g., Mele & Robb 1998, 2003 and Pereboom 2001, 2005) have suggested that there *can* be cases in which a person's action is not determined yet the intervener eliminates all of the person's genuine or robust alternative possibilities. In this chapter, I extend the original dilemma defense by arguing that these new lines of defense of Frankfurt's conclusion are unpersuasive and fall prey to the original dilemma. I thus conclude that the Frankfurt cases do not show that PAP is false.

Chapter three – 'Raising the Responsibility Question.'

Recently PAP-defenders have developed a different line of response to the Frankfurt cases. They have begun to question the very intuition that Jones should be thought to act freely and responsibly in the first place. They have, as I put it, raised the responsibility question with respect to Jones' blameworthiness. In his original 1969 paper, Frankfurt argues that since the fact that Jones could not have done otherwise is irrelevant to causally explaining Jones' behavior – he would, it seems, have done the same thing for the same reasons even if Black had not been present and he *could* have done otherwise – then it would be gratuitous to assign this fact any weight in the assessment of his moral responsibility. In the first part of the chapter, I critically examine this argument.

In the second part of the chapter, I assess why PAP seems so intuitive. This is important in the context of raising the responsibility question for some PAP-adherents have suggested that Jones should not be thought to act freely and responsibly, in the Frankfurt case, precisely *because* he could not have done otherwise. But for such a move to be plausible, it would be helpful to have an account of PAP's plausibility, to have a justification for PAP itself. I outline a recent argument for PAP's truth by Widerker (2000, 2005) which rests on the claim that unless there is a good answer to the question 'What should Jones have done instead?' then Jones cannot be blameworthy for what he does. I end the chapter by defending this argument against some recent criticisms by Fischer and McKenna.

Chapter four – 'Frankfurt's Indirect Attack on PAP – The Hierarchical Account.'

In chapter four, I turn to Frankfurt's *indirect* attack on PAP developed in the second of his two important papers, 'Freedom of the will and the concept of a person' (1971). In this indirect attack, he argues that there is an alternative conception of freedom different from the freedom to do otherwise. He claims that this alternative conception which is compatible with determinism does *not* rely on alternative possibilities yet *is* sufficient to capture the freedom required for moral responsibility. If Frankfurt is right, then PAP is false.

In the main part of the chapter, I outline and discuss Frankfurt's condition of the freedom required for moral responsibility which turns on distinguishing between higher-order and lower-order desires. Specifically, Frankfurt argues that a person acts with the freedom required for moral responsibility if and only if he acts from a desire that is his will – that is, is his motivating desire – because it was the will he wanted. I try to place Frankurt's account in a broader context by showing that the reason Frankfurt distinguishes between higher-order and lower-order desires is to explain how some of a person's desires can be more truly 'his' than others. After outlining the theoretical background on which his account rests, I end the chapter by looking at Frankfurt's neglected example of the *willing* drug addict, a case that throws further light on Frankfurt's theory.

Chapter five – 'Evaluating Frankfurt's Hierarchical Account.'

Having offered what I think is the most plausible interpretation of Frankfurt's positive account of the freedom required for moral responsibility, I then turn in chapter

five to evaluate it. I focus on two criticisms. First, I assess whether a person must identify himself with his motivating desire in order to be morally responsible as Frankfurt suggests. I focus in particular on apparent counterexamples that suggest that identification is not necessary, though I argue that Frankfurt and his defenders have the apparatus to avoid these alleged counterexamples.

Assuming that identification is necessary for responsibility, the second criticism I look at focuses on whether Frankfurt's condition is *sufficient* to capture the way in which a morally responsible agent identifies himself with his motivating desire. I argue that this criticism is much more troubling for Frankfurt's account than the first one. Criticisms that Frankfurt's condition is not sufficient for identification are not rare, but I try to break new ground by arguing that his account is not subject to one regress difficulty, as is commonly argued, but to two independent regress problems. I assess Frankfurt's responses over the years to this sort of criticism.

Chapter six – 'Undermining New Compatibilism I – Manipulation and Causal Responsibility.'

In the final two chapters of the dissertation, I turn away from Frankfurt's specific account and look more generally at compatibilist views that try to capture moral responsibility's freedom without reference to alternative possibilities, so called *new compatibilist* views. Besides Frankfurt's account, I consider the new compatibilist views of John Martin Fischer and Mark Ravizza (1998), T. M. Scanlon (1998), Angela Smith (2005, 2008), R. Jay Wallace (1994), and Gary Watson (1975).

Generally speaking, these new compatibilist views that do not make use of the freedom to do otherwise can be divided into two kinds depending on whether they emphasize self-expression or responsiveness to reasons as the key freedom-relevant feature of moral responsibility. On the self-expression model, people act with the freedom required for moral responsibility to the extent that their actions reflect the parts of their selves that are most fundamental to who they really are, as people. On the responsiveness to reasons picture, people act with the freedom required for moral responsibility to the extent that they regulate their behavior by moral reasons.

In this chapter, I outline and evaluate two strategies by which PAP might be defended in the face of these new compatibilist conditions. The first involves manipulation cases, while the second exploits the apparent difference between causal responsibility and moral responsibility. Despite my sympathies with the aims of these arguments, however, I conclude that neither forms a decisive strike against the sufficiency of the new compatibilist conditions.

Chapter seven – 'Undermining New Compatibilism II – Blame, Demands, and Authorship.'

In this final chapter, I develop an argument for PAP's truth based on some recent work by Widerker (2005). It draws on the apparent link between moral blame on the one hand and moral demands on the other. I suggest the following principle:

> *The Principle of Alternative Demands (PAD):* An agent *S* is morally blameworthy for doing *A* only if in the circumstances it would be reasonable for those in a position to do so to demand that *S* not have done *A*.

PAD supports the claim that blame requires alternative possibilities because in order for it to be reasonable for such a demand to be made it must be the case that the individual could have done otherwise. For how can it be reasonable to make demands of people if it is not within their power to meet those demands?

If sound, this argument shows that the new compatibilist conditions are not sufficient to capture the freedom required for moral responsibility, at least as it applies to moral blame. But the argument has broader implications. It shows that a traditional assumption about the relationship between freedom and moral responsibility is false. According to this traditional assumption, the issue of a person's freedom as it pertains to his moral responsibility is simply the issue of determining the kind of freedom needed for him to bring about, or 'author,' his behavior. Yet this argument shows that there is *more* to delineating a person's freedom when assessing his moral responsibility than simply determining this. In addition, we must *also* ask whether the individual acted with sufficient freedom for it to be reasonable to demand that he not have acted as he did and, instead, have acted as morality requires.

Chapter Two: Frankfurt's Direct Attack on PAP – The Frankfurt Cases.

In this chapter, I defend the truth of PAP against apparent counterexamples to it. These alleged counterexamples, known as 'Frankfurt cases' from Frankfurt's (1969) well-known presentation of them, have convinced many that moral responsibility does not require the freedom to do otherwise. Here, I develop and defend one particular objection to the Frankfurt cases known as the *dilemma defense* of PAP. According to this line of defense, these examples fall prey to a dilemma, both horns of which undermine their cogency.

2.1 The Frankfurt cases.

Here is an example of a Frankfurt case.[7] Imagine a person, Jones, decides to lie to save embarrassing himself, despite knowing that it is morally wrong for him to do this. Unbeknownst to Jones, he is unable to do otherwise because of the presence of Black. Black is a counterfactual intervener, someone who could and would intervene to make Jones lie, if Jones had not decided to lie 'on his own.' But, as Frankfurt (1969) argues in his original article, 'Alternate possibilities and moral responsibility,' since Black's presence has no effect on what Jones did – Jones would, it seems, have done the same thing for the same reasons even if Black had not been present and Jones *could* have done otherwise – it seems right to think that Jones is blameworthy, and hence morally responsible, for lying despite lacking the freedom to do otherwise. Hence, PAP is false.

---

7 This case is similar to one outlined by Haji & McKenna (2004).

As Frankfurt (1969) himself acknowledges, though, the example is underdescribed in its original form (p. 835, footnote 3). This is because we can ask how Black can have the counterfactual power to ensure that Jones can only decide how Black wants him to. In the case as it is initially presented, should Jones, in the counterfactual scenario, decide *not* to lie, then it seems that Black can only intervene and 'make' Jones decide to lie *after* Jones has begun to decide not to lie. There is no way for Black to know that he needs to intervene until that point. But the fact that Black can only intervene at this point in the counterfactual scenario, after Jones has begun to decide not to lie, seems to ensure that, in the actual circumstances, Jones has alternative possibilities at the moment of choice, lying or not lying.

The examples have been subsequently refined to avoid this problem by using the idea of a 'sign' Jones displays just prior to his deciding to lie. We can imagine that Jones is of a nervous disposition and just before any instance in which he decides to do something that he knows to be morally wrong – in this case, deciding to lie to save embarrassment – he involuntarily twitches. Black knows this fact about Jones and can use it to ensure that Jones has no choice but to decide to lie. This is because Black knows that the circumstances are such that should Jones be about to decide to lie, he will involuntarily twitch just prior to making the decision. But should Jones be about to decide *not* to lie, he will *not* twitch before the decision. With this knowledge, Black notices, in the actual circumstances, that Jones twitches and so Black stays hidden, knowing that Jones will decide just as Black wants him to. Had he not detected a twitch, then Black could have prevented Jones from doing anything other than deciding to lie

because he would have intervened and 'made' Jones decide to lie before Jones could have decided *not* to lie.

These prior sign examples ensure that Black, the fail-safe mechanism, has sufficient counterfactual power to rule out Jones' alternative possibilities prior to the beginning of a freely willed action. However, some have suggested that even in these new circumstances, Jones still would have an alternative possibility open to him that Black cannot eliminate – the alternative, for instance, *not* to twitch. But even if this were true, the availability of this alternative, being an involuntary twitch, would not be something that would be within Jones' voluntary control. Furthermore, it seems intuitive to think that if an alternative possibility is to be cited as partly explaining why a person is morally responsible for what he does, it must have a certain character. Namely, it cannot be such that it would be outside the scope of the person's voluntary control. As John Martin Fischer (Fischer, Kane, Pereboom & Vargas 2007) puts this intuition:

> … just as it is not enough to secure moral responsibility that a different choice could have *randomly* happened, it does not seem to be enough to secure moral responsibility that … [the absence of the prior sign] could have been exhibited *involuntarily*. … How could something as important as moral responsibility come from something so thin – and something entirely involuntary? (pp. 58-59).

Descriptions of this intuition give rise to the following principle:

The Robustness Principle:

If an alternative possibility is to be legitimately cited as partly explaining why a person is morally responsible for his action, then that alternative must be one that is within the person's voluntary control.

In what follows I take the robustness principle to be an important restriction that PAP defenders should abide by in responding to Frankfurt cases.

2.2 The Dilemma defense.

Assuming that defenders of PAP should adhere to the robustness principle, how else might they respond to Frankfurt's example besides pointing to the presence of involuntary alternatives? There is an additional problem, I believe, with Frankfurt's example and all other so called Frankfurt examples. They all fall prey to a dilemma, both horns of which undermine their cogency. According to the dilemma, the cases either beg the question against the incompatibilist or they fail to describe circumstances in which the fail-safe mechanism (Black, in the original example) eliminates all of the person's robust alternative possibilities. (By speaking of his 'robust alternative possibilities' I mean those that are within the scope of his voluntary control for it is only the existence of these alternatives I am assuming, and not any involuntary possibilities, that could bear on the question of his moral responsibility.)

The dilemma turns on whether or not the involuntary twitch Jones displays prior to his decision to lie is deterministically related to his subsequent decision to lie. On one horn of the dilemma – the *indeterminism horn* – if the twitch is *not* deterministically related to Jones' subsequent decision, then Black will not have an exact basis for knowing how Jones will decide. For if the twitch is not deterministically related to Jones' subsequent decision, then upon displaying the sign, it will still be open what decision (if any) Jones will make. In such circumstances, Black will have to wait until *after* Jones has begun to decide one way or the other, to lie or not to lie, in order to know

whether or not he needs to intervene. However, his waiting ensures that, at the moment of choice, Jones could have done otherwise. If Black were to intervene *before* Jones has begun to make a decision one way or the other, then he, but not Jones, would be morally responsible for the decision.

On the other horn of the dilemma – the *determinism horn* – if the twitch *is* deterministically related to the decision itself, then while Black would have an exact basis for knowing which way Jones will decide, it would beg the question against incompatibilists to argue that Jones could be morally responsible under such circumstances despite lacking alternative possibilities. For if Jones' decision has a deterministic cause, then incompatibilists cannot, from the outset, think that Jones is morally responsible for his decision.

This line of defense against the Frankfurt cases, known as the 'dilemma defense,' was originally developed by Robert Kane (1985) and subsequently taken up by Carl Ginet (1996) and David Widerker (1995). Since then, those wanting to reject PAP have responded in two ways. Some (e.g., Fischer 1999, 2006, and Haji & McKenna 2004) have argued that a causally determined Frankfurt case *can* be used to show that PAP is false without begging the question against the incompatibilist. Others (e.g., Mele & Robb 1998, 2003 and Pereboom 2001, 2005) have suggested that there *can* be cases in which the prior sign is not deterministically related to the person's subsequent action yet the fail-safe mechanism eliminates all of the person's robust alternative possibilities. In what follows, I extend the original dilemma defense by arguing that these new lines of response to it are unpersuasive and fall prey to the original dilemma. I thus conclude that the Frankfurt cases, even in their recently revised form, do not show that PAP is false.

2.3 The Determinism horn.

With respect to the determinism horn of the dilemma defense, Haji and McKenna (2004) have recently argued that the question-begging charge, and the dialectic surrounding it, is subtler than has been appreciated. They distinguish between two different audiences to whom Frankfurt's argument can be directed. These correspond to, what Haji and McKenna call, two different "dialectical contexts" (p. 302). They believe that these two different dialectical contexts give rise to two different readings of the question-begging charge. On the one hand, they admit that the charge may be pertinent in one of the contexts, what they call the *broad* context in which Frankfurt is trying to persuade the *committed incompatibilist* that PAP is false. On the other hand, they argue that there is reason to believe that the question-begging charge is not sound in the other context, the *narrow* context in which Frankfurt intends to convince an audience that is *undecided about compatibilism or incompatibilism* that PAP is false.[8]

Turning to the *broad* dialectical context in which Frankfurt intends to convince a committed incompatibilist that moral responsibility does not require the freedom to do otherwise, Haji and McKenna agree that "the Dilemma Defender is entitled to the complaint that the first horn of the dilemma [assuming determinism] is question-begging" (p. 313). This is because if the Frankfurt cases are to persuade a committed incompatibilist that PAP is false, then they cannot assume the truth of a condition, the condition of determinism, that the incompatibilist believes rules out Jones' being

---

8 I am setting aside another possible dialectical context in which Frankfurt means to persuade a *traditional compatibilist* that PAP is false. This is a compatibilist who believes that moral responsibility requires the freedom to do otherwise but understands this freedom in compatibilist-friendly terms.

blameworthy for lying in the first place. In the broad dialectical context, in which Frankfurt is trying to convince a committed incompatibilist that PAP is false, "it should be clear that the Dilemma Defender's appeal to the first horn *is unimpeachable*; it is simply a condition of the debate that the Frankfurt Defender offer a case that does not assume determinism" (p. 304). If Frankfurt and his defenders want to work within the broad dialectical context of trying to convince committed incompatibilists that PAP is false, then they must – Haji and McKenna agree – offer an example that does not assume determinism (as, for instance, Mele and Robb [1998] and Pereboom [2001, 2005] do, examples I discuss later in this chapter).[9]

But turning to the *narrow* dialectical context, in which Frankfurt means to persuade an audience that is undecided about compatibilism or incompatibilism that PAP is false, Haji and McKenna insist the situation is different. In this context, those defending the dilemma cannot legitimately assert that a determined Frankfurt case is

---

9 However, Haji and McKenna argue that not *all* incompatibilists are entitled to assert the broad interpretation of the first horn. An incompatibilist whose *sole* reason for thinking that determinism is incompatible with moral responsibility is that determinism eliminates alternative possibilities would be behaving in "bad faith" (p. 312), Haji and McKenna argue, should he claim that assuming determinism begs the question simply because a deterministic relation obtains. This is because "by demanding that Frankfurt begin with a case of a libertarian free agent … he is demanding that Frankfurt provide a case in which any freely willing and morally responsible agent does have alternative possibilities. This of course is just what Frankfurt wants to disprove" (p. 313). Haji and McKenna argue that if this sort of incompatibilist wants to rely on the dilemma defense, he "must accept the conditions of the debate established by the narrow dialectical context" (p. 312). By contrast, Haji and McKenna claim that those incompatibilists who believe that determinism is incompatible with freedom and moral responsibility for *other* reasons than determinism's apparent elimination of alternative possibilities – those who believe, for instance, that determinism is incompatible with a person being the genuine source of his actions, something that is required for moral responsibility – *can*, in good faith, assert the broad interpretation of the determinism horn of the dilemma defense. I set this complication aside in what follows.

question-begging simply because a deterministic relation obtains simpliciter. An audience that is undecided about the truth of compatibilism or incompatibilism should not, on the face of it, find the fact that Jones' decision is determined by his prior sign objectionable with respect to the issue of his blameworthiness. This is because they have no commitments about whether determinism is compatible or incompatible with moral responsibility.

However, Haji and McKenna argue that there is another version of the question-begging charge potentially available to those defending the dilemma, one that does not rely on the fact that what is allegedly question-begging in the cases is that a deterministic relation obtains simpliciter. On this alternative picture, the question-begging feature of causally determined Frankfurt cases is not simply that determinism is assumed but rather than "*the deterministic relation expunges alternative possibilities*" (p. 303). That is, the dilemma defender could argue that assuming a determined relationship between the sign and Jones' decision begs the question against the incompatibilist for the following reason. If determinism is incompatible with the freedom to do otherwise as incompatibilists maintain, then should the undecided audience grant that it is possible for Jones to be blameworthy prior to Black's introduction into the example, they will have shown themselves to have granted Frankfurt's point, that moral responsibility does not require the freedom to do otherwise, from the beginning. As Haji and McKenna put it:

> If an undecided audience grants that such an agent [in a deterministic setting] is even prima facie free and morally responsible without the machinery [i.e., Black's presence], as incompatibilists might see it, this audience will have been duped into granting Frankfurt's point [that PAP is false] at the outset (p. 306).

So, on this version of the question-begging charge, "to insure an important question is not begged," Frankfurt and his defenders "need a case in which the *only* manner of ruling out alternative possibilities is by way of the counterfactual intervener" (p. 309).

However, Haji and McKenna argue that this demand is too strong. While they admit that it would be question-begging against the incompatibilist if Frankfurt were to use a determined example to try to persuade *committed incompatibilists* that PAP is false, they do not believe that it is question-begging to use such a case to persuade an *undecided audience* of this conclusion, even if the question-begging charge centers upon determinism's elimination of Jones' alternative possibilities.[10]

2.4 Criticizing Haji and McKenna.

Haji and McKenna (2004) argue that a determined Frankfurt case can be used without impropriety to persuade an audience who are undecided about compatibilism or incompatibilism that PAP is false because even if determinism eliminates Jones' freedom to do otherwise as incompatibilists claim, there are reasons to think that this would be irrelevant to the question of Jones' blameworthiness. They argue as follows:

(1) The counterfactual intervener, Black, and causal determinism are individually sufficient to eliminate Jones' freedom to do otherwise.

10 Though they agree that "the Frankfurt Defender's case would be more convincing – there would be less to contest – if it were beyond dispute that an agent's alternatives were ruled out *only* by virtue of the counterfactual intervener and not by virtue of the truth of determinism" (pp. 309-310).

(2) Black's elimination of Jones' freedom to do otherwise is irrelevant to the question of his blameworthiness.

(3) Therefore, it is plausible to think that determinism's elimination of Jones' freedom to do otherwise is similarly irrelevant to the question of his blameworthiness.[11]

In other words, even if determinism eliminates a person's freedom to do otherwise as incompatibilists contend, so does Black, the counterfactual intervener. Furthermore, since Black's elimination of Jones' alternative possibilities is irrelevant to the question of his moral responsibility, then determinism's elimination of his alternative possibilities should similarly be irrelevant to the issue of Jones' moral responsibility.

However, in my paper, 'New distinctions, same troubles: a reply to Haji and McKenna' (2005), I argue that this argument is not sound. Premise (1) is false. Causal determinism and the counterfactual intervener are *not* individually sufficient to eliminate Jones' freedom to do otherwise. Only determinism is, in itself, sufficient for eliminating

---

11 Haji and McKenna (2004) write that "even if determinism rules out alternative possibilities," as incompatibilists maintain, "so does the presence of Frankfurt's counterfactual machinery." And since all agree that "the latter way of ruling out alternatives is irrelevant to an evaluation of an agent's freedom and responsibility" – Jones would, it seems, have done the same thing for the same reasons even if Black had not been present and he *could* have done otherwise – this fact "provides plausible grounds for the suggestion that the former is as well" (p. 313). Fischer (2006) makes a similar sort of argument. He writes: "I have claimed that consideration of the examples of Jones (a typical Frankfurt-type case) should first elicit the intuition that the fact that there is a fail-safe device present that would intervene in the counterfactual scenario is *irrelevant* to the grounding of Jones' moral responsibility. My contention is that this then *suggests* that even if Jones had *no* alternative possibilities at all, this would be irrelevant to the grounding of his moral responsibility. It would then follow that in a causally deterministic world, in which it is assumed that Jones has no alternative possibilities at all, his lack of alternative possibilities would be irrelevant to the grounding of his moral responsibility" (p. 198).

his alternative possibilities. The counterfactual intervener, Black, can only eliminate Jones' freedom to do otherwise if we assume that the relationship between Jones' prior sign and his subsequent decision to lie is determined.

To see that this is so, return to the original dilemma. In order for Black to be able to eliminate Jones' freedom to do otherwise, there must be a determined relationship between Jones' prior twitch and his subsequent decision. If the twitch is not deterministically related to his subsequent decision, then Black does not have an exact basis for prediction. He will have to wait until after Jones has begun to decide one way or the other, to lie or not to lie, in order to know whether or not he needs to intervene. His waiting, however, ensures that, at the moment of choice, Jones could have done otherwise.

So, for the counterfactual intervener to eliminate Jones' freedom to do otherwise, there *must* be a determined relationship between Jones' prior sign and his subsequent decision. Since premise (1) of the argument is false, then the conclusion, (3), that it is plausible to think that determinism's elimination of Jones' freedom to do otherwise is irrelevant to the question of his blameworthiness, does not follow. We have not, therefore, been given good reason to think that determinism's elimination of Jones' freedom to do otherwise is irrelevant to the question of his moral responsibility. In fact, incompatibilists will insist that the way determinism eliminates Jones' alternative possibilities prior to Black's introduction into the example is *very* relevant to the issue of Jones' blameworthiness.

2.5 Haji and McKenna's response.

Haji and McKenna respond to my argument in their paper, 'Defending Frankfurt's argument in deterministic contexts: a reply to Palmer' (2006). They summarize my contention as to why premise (1) of their argument is false as follows:

> In a prior-sign Frankfurt case, the presence of the counterfactual intervener (Black), *only* in conjunction with an independently sufficient condition for Jones' choosing to act as Black desires (the condition that Jones' choice is causally determined), ensures that Jones has no alternatives (pp. 368-369).[12]

They then sketch a line of reply to my criticism:

> What Palmer's objection requires, really, is the stronger contention that it is not possible for there to be a case in which the counterfactual intervener eliminates alternatives without the assumption of determinism. For if there were such a case, the claim that the counterfactual intervener *qua* counterfactual intervener is sufficient to rule out alternative possibilities would be vindicated. And then surely we could just wed one of those cases to a deterministic context (p. 369).

In other words, Haji and McKenna want to avoid my criticism by two steps of reasoning:

> *Step one:* Outline a Frankfurt case in which the counterfactual intervener eliminates the person's freedom to do otherwise without needing to rely on a deterministic relationship between the individual's prior sign and his subsequent decision.

---

12 Here, and in what follows, I've altered the names of the individuals Haji and McKenna cite to fit with the presentation of my case.

Step two: Modify this case by making the relationship between the individual's prior sign and his subsequent decision determined. Such a modification – the 'adding' of determinism – would not be required for the counterfactual intervener to have the power to eliminate the person's freedom to do otherwise. This is because if determinism were not true in the example, the counterfactual intervener would still have sufficient counterfactual power.

Let us suppose for the sake of discussion that if this two-step strategy works, Haji and McKenna will have shown that premise (1), the claim that the counterfactual intervener and causal determinism are individually sufficient to eliminate Jones' freedom to do otherwise, is true. If they show this premise to be true, then they will be able to conclude that there *are* sound reasons to think that determinism's elimination of Jones' freedom to do otherwise would be irrelevant to assessing his blameworthiness. Moreover, if *this* were true, then incompatibilists could not legitimately complain that people would be duped into believing Frankfurt's conclusion if they granted that Jones could be morally responsible for his decision prior to Black's introduction into the example. This is because even if incompatibilists are right to think that determinism would eliminate Jones' alternative possibilities, they would be wrong to insist that this fact is relevant to an assessment of his blameworthiness.

The controversial part of Haji and McKenna's reply is the first step, the claim that there *can* be Frankfurt cases in which the person's prior sign is not deterministically related to his subsequent decision, yet the fail-safe mechanism eliminates all of the person's robust alternative possibilities. Two recent examples that have convinced many that such cases are possible have been developed by Mele and Robb (1998, 2003) and

Pereboom (2001, 2005). (Haji and McKenna make use of Pereboom's example to illustrate their contention that there can be such cases.) In the final sections of the paper, I criticize these two new examples. If I can undermine these cases, then I will have cast doubt on the soundness of Haji and McKenna's first step in their reply to my criticism.

The examples by Mele and Robb and by Pereboom are of considerable interest beyond whether or not Haji and McKenna can make use of them. After all, if it could be demonstrated that it is possible for Black to eliminate Jones' robust alternative possibilities without needing to rely on a deterministic relationship between Jones' prior sign and his subsequent decision, then the issue of whether or not the deterministic relation begs the question against the incompatibilist would be moot. Those wanting to reject PAP by using the Frankfurt cases would have successfully avoided the dilemma defense on the *other* horn of the dilemma, the indeterminism horn. With this in mind, let me turn to assess the soundness of the indeterminism horn of the dilemma defense.

2.6 The Indeterminism horn I: Pereboom.

According to the indeterminism horn of the dilemma defense, should the prior sign not be deterministically related to the individual's subsequent decision, then the counterfactual intervener will be unable to eliminate all of the person's robust alternative possibilities. In what follows, I outline and criticize recent examples by Pereboom and Mele and Robb that attempt to avoid this criticism.

Here is Perboom's (2005) considered presentation of his case, 'Tax evasion,' one he claims avoids the indeterminism horn of the dilemma:

> Joe is considering whether to claim a tax deduction for the substantial local registration fee that he paid when he bought a house. He knows that claiming the deduction is illegal, that he probably won't be caught, and that if he is, he can convincingly plead ignorance. Suppose he has a very powerful but not always overriding desire to advance his self-interest regardless of the cost to others, and no matter whether advancing his self-interest involves illegal activity. Crucially, his psychology is such that the only way that in this situation he could fail to choose to evade taxes is for moral reasons. His psychology is not, for example, such that he could fail to choose to evade taxes for no reason or simply on a whim. In addition, it is causally necessary for his failing to choose to evade taxes in this situation that he attain a certain level of attentiveness to these moral reasons. He can secure this level of attentiveness voluntarily. However, his attaining this level of attentiveness is not causally sufficient for his failing to choose to evade taxes. If he were to attain this level of attentiveness, Joe could, with his libertarian free will, either choose to evade taxes or refrain from so choosing (without the intervener's device in place). More generally, Joe is a libertarian free agent. But to ensure that he chooses to evade taxes, a neuroscientist now implants a device, which, were it to sense the requisite level of attentiveness, would electronically stimulate his brain so that he would choose to evade taxes. In actual fact, he does not attain this level of attentiveness, and he chooses to evade taxes while the device remains idle (pp. 231-232).

Pereboom claims that the neuroscientist's elimination of Joe's alternative possibilities should not negate Joe's moral responsibility for his decision because Joe

would have done the same thing for the same reasons even if the neuroscientist had not been present and he could have done otherwise. Pereboom accepts the robustness principle as a restriction that PAP defenders should abide by in responding to Frankfurt cases. He gives an account of robustness, arguing that an alternative possibility is robust, and so one that be legitimately cited as partly explaining why a person is morally responsible for his action, if and only if the individual "could have willed something other than what she actually willed such that she understood that by willing it she would thereby have been precluded from the moral responsibility she actually has for the action" (p. 230).

Pereboom admits that Joe does have *an* alternative possibility open to him – the possibility to voluntarily achieve the level of attentiveness to moral reasons – but he argues that this is not a *robust* alternative, and so no help to the PAP defender. He offers two reasons why this possibility is not robust. First, Joe does not "believe that if he had achieved the requisite level of attentiveness he would thereby have been precluded from responsibility for deciding to evade taxes" (p. 232). The reason for this is because his acquiring the requisite level of attentiveness is only a *necessary*, and not sufficient, condition for his deciding *not* to evade taxes (and he knows this fact). It is still open to him, on achieving the requisite level of attentiveness, to decide *to* evade taxes. Second, "Joe does not know enough to understand that voluntarily achieving the requisite attentiveness would preclude him from responsibility for choosing to evade taxes" (p. 232). Given that the neuroscientist's involvement is unknown to Joe, "Joe doesn't understand that the intervention would then take place [were he to achieve the requisite level of attentiveness], or that as a consequence of this intervention he would be

precluded from responsibility for choosing to evade taxes" (p. 233). His ignorance of the neuroscientist's involvements ensures that he could not know that by doing something else – i.e. by achieving the requisite attentiveness – he would (because of the neuroscientist's subsequent intervention) be precluded from responsibility for choosing to evade taxes (because the neuroscientist would intervene and 'make' Joe make that choice).

According to the dilemma that plagued Frankfurt's original alleged counterexample to PAP, recall, if the target action is not deterministically related to the individual's prior sign, then while not question-begging against the incompatibilist, the individual will have alternative possibilities open to him because the intervener will have to wait to see – that is, wait until the person has begun to act one way or the other – in order to know whether he must intervene. Pereboom believes 'Tax Evasion' avoids this problem because, while Joe's action is causally indetermined, he does not have open to him an alternative possibility that meets the robustness condition. Joe knows that his voluntarily achieving the level of moral attentiveness is compatible with his still choosing to evade taxes (so it is not a robust alternative possibility to his evading taxes) and he also does *not* know that, by doing this, he would – because of the intervener's intervention – have avoided responsibility for his action (thus failing to meet the condition for robustness that he should know that by willing the alternative action he would thereby – because of the intervener's intervention – be precluded from the moral responsibility he actually has).

In what follows, I argue that by paying closer attention to the details of his example, we see that Pereboom's case fails to avoid the indeterminism horn of the

dilemma. Specifically, I claim that by paying attention to the precise time at which Joe's action is to be performed, we see that Joe *does* – contrary to Pereboom's claim – have a robust alternative open to him on Pereboom's definition of robustness, an alternative the neuroscientist does not have it within his power to eliminate. Thus, 'Tax Evasion' does not constitute a counterexample to PAP.

2.7 Criticizing Pereboom.

Call the time at which Joe makes the decision to evade taxes *t1*, label the decision *A*, and call the requisite level of attentiveness to moral reasons (necessary but not sufficient for Joe to fail to decide *A*) *B*. Suppose that Joe had made *B* occur at *t1*, rather than – as he actually did – decide *A*. Now, we can agree with Pereboom that Joe's making *B* occur is only a necessary and not sufficient condition for his failing to decide to evade taxes *simpliciter*. However, his making *B* occur *is* sufficient for his *failing to decide A at t1*, because, at *t1*, Joe is making *B* occur instead.

The neuroscientist's device could not have intervened, and made Joe decide *A*, until some time *t2* later than *t1* (that is, some time *t2* that begins later than *t1*). This is because since Joe's making *B* occur at *t1* is indeterministic, it must be open up until *t1* that Joe could decide *A* at *t1* rather than make *B* occur. From this it follows that there exist no facts that entail that *B* occurs until a time after *t1* has begun, because up until *t1* all the facts are consistent with Joe's deciding *A* at *t1* instead. If Joe's making *B* occur at *t1* is indeterministic, then the facts entailing that *B* occurs at *t1* only begin to obtain – they only come into *being* – when *B*'s occurring has begun to happen. (On the other hand, if Joe's making *B* occur at *t1* were *deterministically* caused, then immediately before *t1*

begins there would exist facts entailing that *B* would occur at *t1*, namely, the fact that the deterministic cause has occurred and the fact that deterministic causal laws entail that such a cause produces *B*.)

Necessarily, the intervening device could detect the occurrence of *B* only at a point when the facts entail that *B* occurs at *t1*. Only then, at *t2*, once it becomes true (the facts entail that) *B* has begun to occur, could the device 'know,' so to speak, that (i) *B* has begun to occur, and so (ii) it should intervene and 'make' Joe decide *A*.[13] The upshot is that Joe had a robust alternative possibility open to him at *t1*, the time he, in fact, decided *A*, to evade taxes. The robust possibility is not that Joe had it open to him, at *t1*, to choose not to evade taxes *simpliciter* – the presence of the device ensures that, *by t2*, Joe will choose *B*. Rather, he had it open to him, at *t1*, not to choose to evade taxes *at t1*.

Ginet (2002) points out that, should the neuroscientist be unable to eliminate this alternative as I have argued, it would be robust according to Pereboom's definition of robustness. This is because should Joe have taken it, and made *B* occur at *t1*, he would then – before the device could 'know' that it must intervene (the device could not know this, as I have argued, until a later time *t2*) – know that by refraining from deciding *A* at that time, he would be avoiding responsibility for doing *A* at that time, *t1*. He would know this simply in virtue of his *not* doing *A* right then – at that time, *t1* – which he would, of course, be aware of. (Recall that Pereboom contends that an alternative

---

13 As Carl Ginet encouraged me to make clear, the point is not that Joe could begin to make *B* occur at *t1* only after a short time after *t1* has begun; Joe's making *B* occur at *t1* is compatible with its being the case that his beginning to make *B* occur coincides with the beginning of *t1*. Rather, the important point is that, even so, it becomes true that *B* has begun to occur – something of *B* has occurred – only at some instant *t2* later than the

possibility is robust if and only if the agent could have willed something other than he did and he knew that by willing it he would have avoided the moral responsibility he, in fact, had for the action he, in fact, performed.)

Summarizing the argument, the set up in 'Tax Evasion' is such that the counterfactual intervener cannot make it the case that the agent, Joe, could not have avoided acting in the way he did at the precise time he did so. The alternative possibility open to him, at the time he acted, the possibility not to decide to evade taxes at *t1*, is – on Pereboom's own definition – a robust alternative possibility.

Pereboom (2005) offers a reply to Ginet's point that the alternative possibility open to Joe is robust. He begins by saying that

> Ginet's contention depends on the claim that given that the mechanism would have made Joe choose to evade taxes after *t1*, he would have been aware at *t1* that he was not making this choice, and he would have understood at *t1* that by not choosing to evade taxes right then, he would not have been responsible for making this choice right then (p. 234).

Then he amends his original example in the following way:

> In the neuroscientist's setup, Joe's making the necessary condition for doing otherwise [his making *B* occur at *t1*] instantaneously renders him unconscious for several minutes (p. 234).

The device's rendering Joe instantly unconscious, should he make *B* occur at *t1* rather than decide *A*, is supposed to ensure that, at *t1*, Joe does not have a robust

---

beginning of *t1*. And, necessarily, the intervening device could detect the occurrence of *B* only at a point when *B* has begun to occur.

alternative possibility open to him. His being unconscious removes any open robust alternative possibility because – supposing he had made *B* occur at *t1* – "at the next instant … Joe would not have been aware that he was not choosing to evade taxes, and he would not [at that instant]… have understood that by not choosing to evade taxes right then he would not have been responsible for choosing to evade taxes right then" (p. 234).

We can suppose, for the sake of discussion, that Pereboom is right that, were Joe rendered instantaneously unconscious by the device should he have made *B* occur at *t1*, he would not have a robust alternative possibility open to him at *t1*. However, the device cannot have the counterfactual power to render Joe unconscious *at t1*, were he to have made *B* occur at *t1*. The device can only render Joe unconscious a short time, *t2*, that begins after *t1* begins and because of this time gap, at *t1*, Joe will still have a robust alternative possibility open to him.

The reason that the device cannot render Joe unconscious at *t1* but only at *t2* is the same reason that the device cannot intervene and 'make' Joe decide *A* at *t1* should he make *B* occur then instead. Since Joe's making *B* occur at *t1* is indeterministic, then the facts entailing that *B* occurs at *t1* only begin to obtain – they only come into *being* – when *B*'s occurring has begun to happen, that is at a short time, *t2*, that begins later than *t1*. Only then, at *t2*, once it becomes true – the facts entail that – *B* has begun to occur, could the device 'know,' so to speak, that (i) *B* has begun to occur, and so (ii) it should intervene and render Joe unconscious. Pereboom's amended example fails to make it the case that Joe does not have a robust alternative possibility open to him because the intervener's device cannot – contrary to Pereboom's claim – have sufficient

counterfactual power to ensure that, should Joe have made *B* occur at *t1*, he would then – at *t1* – be rendered unconscious.

I now turn to examine Mele and Robb's Frankfurt case.

2.8 The Indeterminism horn II: Mele and Robb.

Mele and Robb (1998) have offered another variant on Frankfurt's original example that they claim avoids the horns of the dilemma defense. Here is their case:

> Our scenario features an agent, Bob, who inhabits a world at which determinism is false … At *t1*, Black initiates a certain deterministic process *P* in Bob's brain with the intention of thereby causing Bob to decide at *t2* (an hour later, say) to steal Ann's car [call this decision *E*]. The process, which is screened off from Bob's consciousness, will deterministically culminate in Bob's deciding at *t2* to steal Ann's car unless he decides on his own at *t2* to steal it or is incapable at *t2* of making a decision (because, for example, he is dead by *t2*). (Black is unaware that it is open to Bob to decide on his own at *t2* to steal the car; he is confident that *P* will cause Bob to decide as he wants Bob to decide) … As it happens, at *t2* Bob decides on his own to steal the car, on the basis of his own indeterministic deliberation about whether to steal it, and his deliberation has no deterministic cause [call this indeterministic deliberation process *x*]. But if he had not just then decided on his own to steal it, *P* would have deterministically issued, at *t2*, in his deciding to steal it. Rest assured that *P* in no way influences the indeterministic decision-making process that actually issues in Bob's decision (pp. 101-102).

What is unique in Mele and Robb's example is that the intervention is not *counterfactual*, as it was in Frankfurt's original case and in Pereboom's example. Rather, in Mele and Robb's example, Black's intervention is *actual*, something that occurs in the actual circumstances. Specifically, Mele and Robb claim that Black's initiation of a mechanism, *P*, inside Bob's brain is such that it would deterministically cause Bob's action at the precise time it actually occurs if Bob were not to perform it 'on his own,' via his indeterministic deliberative process, *x*. In a forthcoming paper, 'On Mele and Robb's indeterministic Frankfurt-style case,' Ginet and I argue that their example is unconvincing. (The basic idea for this criticism is found in Ginet [2003]). Let me explain our argument.

2.9 Criticizing Mele and Robb.

Vital to Mele and Robb's example is the idea that Bob's indeterministic process *x* blocks or preempts the deterministic process *P* from causing *E* at *t2*. With this in mind, we can ask what fact is it about *x* that is supposed to explain how *x* blocks or preempts *P* from causing *E*? In a later article, Mele and Robb (2003: 137, footnote 11) claim that it is *x*'s *causing E* at *t2* that blocks/preempts *P* from causing *E* at *t2*. However, if what allegedly makes it the case that *x* blocks *P* from causing *E* at *t2* is *x*'s *causing E* at *t2*, then Mele and Robb's example is in trouble. This is because, in the counterfactual scenario in which *P* but not *x* causes *E*, *P* can only cause *E* at an instant *t3* later than *t2*. So, if it is *x*'s causing *E* at *t2* that blocks/preempts *P* from causing *E* at *t2*, then it cannot be the case that, if *x* had not caused *E*, then *P* would have caused *E* at precisely the same time. But then Mele and Robb's case cannot be a counterexample to PAP. For a

counterexample to work, it needs to be the case that at the precise time *t2* when he decided to steal the car, Bob could not have *then* done other than decide to steal it. That is, it needs to be the case that if *x* had not indeterministically caused that decision then, *P* would have caused it precisely then. But in Mele and Robb's example, this is not the case.

Why, in the counterfactual scenario, can't *P* cause *E* at *t2* rather than only at a later instant, *t3*? Borrowing from the same general line of reasoning that undermined Pereboom's example, the fact that *P* can only cause *E* at *t3*, rather than at *t2*, follows from the fact that since *x*'s causing *E* is not deterministically caused, it must be open up until *t2* that *x* causes something else or nothing at all at *t2*. From this it follows that there are no facts entailing that *x* causes *E* until an instant after *t2*, because up until that instant the facts are consistent with *x*'s causing something else or nothing at all at *t2*. (On the other hand, if *E* at *t2* were *deterministically* caused, then immediately before *t2* there would be facts entailing that *E* would be caused to occur at *t2*, namely, the fact that the deterministic cause has occurred and the fact that deterministic causal laws entail that such a cause directly produces *E*.)

This criticism is similar, of course, to the original dilemma defense made in response to Frankfurt's original example. According to the dilemma, unless the sign Jones displays prior to his decision is deterministically related to his subsequent decision, then Black will not have an exact basis for knowing what Jones will do. He will have to wait until *after* Jones has begun to decide one way or the other to know whether or not he needs to intervene. Black must wait because Jones' twitching (in the original example) is

not deterministically related to his decision. His waiting, however, ensures that Jones could not have done otherwise.

Making the operation of the fail-safe mechanism *actual* and not merely counterfactual (as Mele and Robb do) does not remove the fundamental problem here. Just as Black has to wait until Bob begins to decide to judge whether to intervene, the process *P* has to wait to see whether the indeterministic process *x* does cause *E* – i.e., has to wait until after *E* has begun to occur – to know whether it can (and must) cause *E*.

Chapter Three: Raising the Responsibility Question

In the previous chapter, I argued that the Frankfurt cases do not show PAP to be false because they fall prey to a dilemma, both horns of which undermine their cogency. Defending or rejecting the soundness of this dilemma has been the main point of controversy between those what want to use Frankfurt's example to reject PAP and those who want to show that his cases do not threaten the principle. However, recently, PAP-defenders have developed a different line of response to the examples. They have begun to question the very claim that Jones, the individual in the Frankfurt case, should be thought to act freely and responsibly in the first place. As I shall put it, PAP-adherent's have *raised the responsibility question* with respect to the cases by challenging the, until now, taken-for-granted claim that when Jones acts in Frankfurt-style circumstances, he does so freely and responsibly. In what follows, I want to evaluate this line of response.

3.1 Raising the responsibility question.

Raising the responsibility question is significant because of the following kind of dialectical situation, aptly described in a recent paper by Michael McKenna (2008a):

> A simplistic interpretation of Frankfurt's argument is that it begins and ends with just that, the production of an example thought to have an intuitively compelling result and in conflict with a principle of moral reasoning [namely, PAP]. But if this were *all* there were to Frankfurt's argument, why should it be so clear that … [the example] should trump the principle? Why not react to the example by saying that, as jarring as it seems, and despite our strong inclination to hold Jones

> blameworthy, the proper judgment is that he is not. For, as the objection might go, it turns out Jones could not do otherwise in acting as he did, and since PAP is such a powerful principle within the arsenal of our moral reasoning, in adhering to PAP, we must resist our intuition regarding the example (p. 772).

McKenna's remarks bring to light the following question: why should we think that Jones is blameworthy for his behavior, given the plausibility of PAP? Indeed, perhaps the correct response is that Jones is not blameworthy for lying precisely *because* he could not have done otherwise.

3.2 The Irrelevance principle.

Rather than appeal to brute intuition in response to this challenge, Frankfurt and his defenders might appeal to some fact about the case, attention to which might invite the thought that, despite PAP's intuitive plausibility, the proper judgment is nonetheless that Jones is morally responsible. We might ask ourselves, on reflection, *what is it* about Jones' situation that invites the thought that he is blameworthy for lying despite the fact that he could not have done otherwise? Frankfurt (1969) suggests an answer to this question in the following passage from his original paper:

> … it would have made no difference, so far as concerns … [Jones'] action or how he came to perform it, if the circumstances that made it impossible for him to avoid performing it had not prevailed. The fact that he could not have done otherwise clearly provides no basis for supposing that he *might* have done otherwise if he had been able to do so. When a fact is in this way irrelevant to the

problem of accounting for the person's action it seems quite gratuitous to assign it any weight in the assessment of his moral responsibility (p. 837).

Frankfurt makes it clear in this passage that what he means by saying that the fact that Jones could not have done otherwise is 'irrelevant to the problem of accounting for the person's action' is that this fact had nothing to do with the "circumstances actually moving him or leading him to do it" (p. 830). The fact that he could not have done otherwise did not "figure at all among the circumstances that actually brought it about that he did what he did" (pp. 836-837). These remarks thus suggest that Frankfurt has in mind something like the following principle:

The Irrelevance Principle (IP):

If a fact is irrelevant to a correct account of the causal explanation of the person's action, then it would be gratuitous to assign it any weight in the assessment of his moral responsibility for his action.

IP can be used to form an argument, on Frankfurt's behalf, that despite the fact that Jones cannot do otherwise, the proper judgment is that he is blameworthy for his decision to lie. Call it

The Irrelevance Argument:

(1) If a fact is irrelevant to a correct account of the causal explanation of the person's action, then it would be gratuitous to assign it any weight in the assessment of his moral responsibility (the irrelevance principle).

(2) The fact that Jones could not have done otherwise is irrelevant to a correct account of the causal explanation of Jones' action.

(3) Therefore, it would be gratuitous to assign it any weight in the assessment of his moral responsibility.

(4) Besides the fact that he could not have done otherwise, there are no other facts that could even prima facie count against Jones' being blameworthy for lying.

(5) Therefore, Jones is blameworthy for lying despite lacking the freedom to do otherwise, and so PAP is false.

The controversial part of this argument, I believe, is premise (1), the irrelevance principle. Is it true that if a fact is irrelevant to a correct account of the causal explanation of a person's action, then it would be gratuitous to assign it any weight in the assessment of his moral responsibility?

3.3 Challenging the irrelevance argument.

David Widerker (2000, 2006) has argued – correctly, to my mind – that the irrelevance principle is false. He claims that "there are intuitive examples that show that sometimes the reason we absolve an agent from blame (or hold him to be blameworthy) for performing a certain act, does not figure at all in the causal explanation of the act" (2006: 178). For the most promising of his counterexamples to IP, consider the following. In the Frankfurt case, Jones decides to lie to save embarrassing himself, despite knowing that it is morally wrong for him to do this. He does not decide to lie *for* the reason that it is morally wrong; he makes the decision *in spite* of knowing this, deciding to lie purely to save public embarrassment. In other words, since Jones decides to lie to avoid embarrassment, we can say that he would have made the same decision whether or not he believed his decision to be immoral.

The fact that Jones knows his decision to lie is morally wrong is irrelevant to a correct account of the causal explanation of his decision and so, according to IP, it would be gratuitous to assign it any weight in the assessment of his moral responsibility. But surely the fact that Jones knows it would be morally wrong to lie yet decides to lie anyway *does* bear on his blameworthiness. A person is more blameworthy, it would seem, if he knows his action is morally wrong and acts anyway in spite of knowing this, than if he does not know it to be morally wrong.

In response to cases like this, Frankfurt (2003) has recently agreed that it would not be gratuitous to assign the fact that Jones knew his action to be morally wrong some weight in the assessment of his blameworthiness. However, he argues that this does not undermine his argument against PAP because "it is considerably less obvious that … this fact plays no role in explaining why … [Jones] acted as he did" (p. 342). Frankfurt argues as follows:

> … to explain an act fully, it is not enough to report merely that he acted for selfish reasons. What counts in the assessment of a person's moral responsibility is not only what causes, reasons, or motives led to his action. It is also important to appreciate what sort of act he thought he was forming. A morally pertinent explanation of what a person has done must include an account of what he believed himself to be doing (pp. 342-343).

With these remarks, Frankfurt agrees that Jones' belief that his decision was immoral is not part of the causal explanation of what he did. However, he suggests that Jones's belief *is* nonetheless part of a 'morally pertinent explanation' of his behavior. Furthermore, it is when a fact is irrelevant to *this* kind of explanation – a morally

pertinent one – and not simply irrelevant to the causal explanation of the action that it would be gratuitous to assign the fact some weight in the assessment of the person's moral responsibility.

3.4 The Revised irrelevance principle.

Frankfurt's remarks thus suggest that he would endorse a revised irrelevance principle:

The Revised Irrelevance Principle (RIP):

If a fact is irrelevant to a correct account of a morally pertinent explanation of a person's action, then it would be gratuitous to assign it any weight in the assessment of his moral responsibility for his action.

This revised principle can also be used on Frankfurt's behalf to construct an argument against the truth of PAP along similar lines to that originating from the original irrelevance principle. Call this new argument

The Revised Irrelevance Argument:

(1') If a fact is irrelevant to a correct account of a morally pertinent explanation of the person's action, then it would be gratuitous to assign it any weight in the assessment of his moral responsibility (the revised irrelevance principle).

(2') The fact that Jones could not have done otherwise is irrelevant to a correct account of a morally pertinent explanation of Jones' action.

(3') Therefore, it would be gratuitous to assign it any weight in the assessment of his moral responsibility.

(4') Besides the fact that he could not have done otherwise, there are no other facts that could even prima facie count against Jones' being blameworthy for lying.

(5') Therefore, Jones is blameworthy for lying despite lacking the freedom to do otherwise, and so PAP is false.

Mirroring the line of reasoning from the original irrelevance argument, absent the presence of any other prima facie reason that could count against his blameworthiness, the proper judgment following from this argument is that Jones is blameworthy for deciding to lie despite lacking the freedom to do otherwise. Hence, PAP is false.

3.5 Challenging the revised irrelevance argument.

I would argue that this revised irrelevance argument is not successful because it is question-begging. The problem is with premise (2'): The fact that Jones could not have done otherwise is irrelevant to a correct account of a morally pertinent explanation of Jones' action.

What is a 'morally pertinent explanation'? On the most natural reading, a morally pertinent explanation of Jones' action would be an explanation that is pertinent to his moral responsibility. However, given this, the argument begs the question against the PAP-adherent. For the only way in which it could be true that the fact that Jones could not have done otherwise is irrelevant to a morally pertinent explanation of his action is by assuming that the fact that Jones could not have done otherwise is irrelevant to an explanation of Jones' moral responsibility. But making this assumption as part of an argument whose conclusion is that PAP is false is question-begging.

3.6 The W-defense.

In response to this question-begging charge, Frankfurt and his defenders might turn the tables and ask why we should think that PAP is so intuitive in the first place. In asking this question, the PAP-rejecters are asking for a justification of PAP itself.

David Widerker (2000, 2005) suggests one line of response to this question, arguing that the reason Jones should be thought blameless for his decision is that his inability to do otherwise ensures that there is no good answer to the question 'What should Jones have done instead?' Unless there is a good answer to this question, according to Widerker, then Jones cannot be blameworthy for his behavior.

Widerker calls this line of argument the 'W-defense,' short for the 'what-should-he-have-done defense'. He writes:

> … since you, Frankfurt, wish to hold him [Jones] blameworthy for his decision … tell me *what, in your opinion, should he have done instead?* Now you cannot claim that he should not have decided [as he did] … since this was something that it was not in Jones' power not to do. Hence, I do not see how you can hold Jones blameworthy for his decision (2000: 191).

Since Jones could not have done otherwise, anyone wanting to hold Jones responsible for his decision cannot provide a satisfactory answer to the question 'What should he have done instead?' They must say either that what he should have done was *not* to have decided as he did (and hence have decided *not* to lie) or that there was *nothing* that he should have done instead. However, these replies are not good answers to the question. The point can be put in terms of fairness. After all, how can it be fair to blame a person for an action if what he *should* have done instead (i.e., *not* to act as he

did) is something that it was not within his power to do? Similarly, how can it be fair to blame a person for an action if there is *nothing else* he should have done instead? If there is no satisfactory answer to what Jones should have done instead, then it seems unfair to think that he is blameworthy for his decision.

Of course, if Jones *could* have done otherwise, then it is plausible to think that a satisfactory answer to what he should have done instead *can* be provided. When asking, in these circumstances, what Jones should have done instead, it is reasonable to reply that he should *not* have decided as he did – he should have decided *not* to lie. Since this was something that it was within his power to do, this answer is legitimate and so no obstruction to his blameworthiness.

Is Widerker's W-defense a good way of defending PAP in the face of the Frankfurt cases? Two philosophers in particular, John Martin Fischer (2006) and McKenna (2005a, 2008a), have raised objections to Widerker's argument. In the remainder of the chapter, I want to defend Widerker's view against these criticisms.

3.7 Responding to the W-defense: Fischer.

Fischer (2006) criticizes Widerker's argument on the grounds that it requires the truth of the maxim 'ought-implies-can,' a maxim he claims is false. However, Widerker (2005) argues that the W-defense, when properly understood, does not rely on this maxim at all. Widerker's argument here turns on the claim that when asking what Jones *should* have done, there are two things that might be being asked: what Jones *ought* to have done, on the one hand, and what it would be *morally reasonable* to have expected him to do, on the other. If what is being asked is what Jones *ought* to have done instead, then

the reply that it is unfair to blame him if what he ought to have done was not something that it was within his power to do *would* seem to require that ought implies can. However, Widerker insists that in asking what Jones should have done, he means to ask what it would be morally reasonable to expect him to have done. With this in mind, it would not require ought implies can to claim that it is unfair to blame Jones if what it would be morally reasonable to expect him to have done was not something that he was free to do.

The idea is that asking what a person morally *ought* to do is to describe his moral *obligations*. Yet, it may not be morally reasonable to expect the person to fulfill his obligations if this would require him to do the impossible and perform an action that he was not free to perform. To illustrate this point, grant that Jones had a moral obligation not to lie – this is what he ought to have done. The issue raised by the W-defense is whether or not it would be morally reasonable to have expected him to meet this obligation. Widerker argues that it would not be morally reasonable to have expected Jones to meet this obligation since this would be to expect him to have done the impossible and do something (i.e., not lie) that he was not free to do.

The issue of the relationship between the W-defense and 'ought-implies-can' is subtle. Moreover, many philosophers are inclined to think that the maxim is true. For these reasons, others have sought to reject Widerker's arguments in ways that do not rely on rejecting this maxim.

3.8 Responding to the W-defense: McKenna.

Once such philosopher is McKenna (2005a) who argues that the best strategy open to the defender of Frankfurt's argument is to:

> … counter Widerker's question with an invitation to consider what the agent *has* done. To Widerker's question, 'What would you have had Jones do?' Frankfurt can reply, 'Look at what Jones has done.' Frankfurt's better answer, I believe, is to admit just for argument's sake that he has no good answer to what he would have had Jones do, but against this intuitively disturbing result, Frankfurt can call attention to what Jones has actually done. The ballgame then comes down to a battle of intuitions (p. 177).

More recently, McKenna (2008a) has called this the 'L-reply,' short for the 'look-what-he-has-done defense'.[14] Its goal is to shift attention *away* from considerations of what alternatives Jones may or may not have had open to him, and *towards* the issue of what Jones actually did. As McKenna acknowledges:

> our conception of the standards of blame and holding responsible *is* tied deeply to a conception of the possible alternatives a morally responsible agent disregards in favor of her settled course of action. Widerker's W-defense forces us to acknowledge that some of our intuitions about alternative possibilities will be resistant to Frankfurt's diagnosis (2005: 177).

However, McKenna argues that his L-reply helps to marshal a different set of compelling intuition about responsibility by inviting us "to fix upon the moral quality of

---

14 Pereboom (Fischer, Kane, Pereboom, & Vargas, 2007), a prominent PAP-rejecter, endorses McKenna's strategy as the best line of reply to the W-defense.

the agent's conduct, to consider what sort of estimation we ought to attach to it" (2008a: 785).

McKenna's reply is interesting, but is it successful? One way to understand the W-defense is as follows:

(1) Someone claiming that Jones deserves blame for lying can be asked 'what should Jones have done instead?'

(2) Unless there is a good answer to this question, then Jones cannot be thought to deserve blame for lying.

(3) There is no good answer to this question in Jones' case.

(4) Therefore, Jones does not deserve blame for lying.

This argument is valid. The question is whether or not it is sound. McKenna thinks that it is unsound. But it is unclear which premise he wants to attack. From the first part of his remarks, he appears to reject premise (3). He suggests that "to Widerker's question, 'What would you have had Jones do?' Frankfurt can reply, 'Look at what Jones has done'" (2005a: 177). One way of interpreting McKenna here is this: he is saying that, contrary to (3), there *is* a satisfactory answer to Widerker's question in Jones' case, namely the answer 'Look at what Jones has done'. And since there is a proper answer to what Jones should have done instead, then this is no obstacle to his being blameworthy. But such a response is unpersuasive. For saying, 'Look what Jones has done,' is not a bona fide reply to the question, 'What should Jones have done instead?' The point is not that the reply is unconvincing. Rather, it is that the remark is not even a candidate reply to that question at *all*.

McKenna's remark is more plausibly construed not as a bona fide reply to Widerker's question, but as a comment made in the light of it. In particular, it is a comment designed to direct intuitions away from those favoring PAP (found in the W-defense) and towards those that center upon what Jones actually did and the moral quality of the will with which he acted. But, if this is how McKenna's argument is to be interpreted, then it is unclear whether he has engaged with Widerker's argument. After suggesting that Frankfurt can reply 'Look at what Jones has done' to Widerker's question, McKenna changes direction by writing that the Frankfurt defender "should admit just for argument's sake that he has no good answer to what he would have had Jones do, but against this intuitively disturbing result, Frankfurt can call attention to what Jones has actually done" (p. 177).

McKenna here seems to be agreeing with premise (3), that there is no satisfactory answer to this question in Jones' case. Which premise then is he challenging? The best interpretation seems to be that he is questioning the truth of premise (2), the claim that unless there is a satisfactory answer to Widerker's question, then Jones cannot be regarded as blameworthy. On this reading, McKenna is claiming that (2) is false: even *if* there is no good answer to the question of what Jones should have done instead, so long as he made his decision 'on his own,' without Black's intervention, then he *can* deserve blame for lying.

But in the light of this claim, the PAP-defender can once again press the responsibility question, and *why* it should be thought right to think Jones blameworthy for a decision to lie he made 'on his own,' given that he could not have done otherwise? The answer given by Frankfurt, McKenna, and others would presumably rest on the some

version of the irrelevance principle, that since Jones' inability to do otherwise is irrelevant to (say) a correct account of the causal explanation of his decision, then it would be gratuitous to assign it any weight in assessing his blameworthiness. However, this principle is false and McKenna's reply to the W-defense cannot succeed by resting on it. So, whether McKenna's L-reply is understood as claiming, on the one hand, that there *is* a good answer to the question of what Jones should have done or, on the other hand, as challenging the claim that unless there is a proper answer to this question, then Jones cannot deserve blame, his reply, to my mind, does not undermine the soundness of Widerker's argument.

Chapter Four: Frankfurt's Indirect Attack on PAP – The Hierarchical Account

In chapters two and three, I argued that direct attacks on the truth of PAP by way of the Frankfurt cases are unconvincing. Despite the arguments of Frankfurt, Fischer, McKenna, Mele, Pereboom, and others, when we look closely at the details of these apparent counterexamples, we see that they do not provide a good reason to abandon the association of moral responsibility with the freedom to do otherwise. In the last four chapters of the dissertation, I want to evaluate the soundness of a different threat to PAP. Aside from direct attacks, the principle has been challenged *indirectly* by compatibilists who argue that there are alternative, compatibilist conceptions of freedom that are different from the freedom to do otherwise. These, so called, *new compatibilists* argue that these alternative conceptions of freedom do *not* rely on alternative possibilities yet *are* sufficient to capture the freedom required for moral responsibility. If they are correct, then PAP would be false. For the freedom pertinent for moral responsibility could then be captured without needing to make use of the freedom to do otherwise.

In his 1971 paper, 'Freedom of the will and the concept of a person,' Frankfurt offered one of the earliest and most influential modern conceptions of freedom not reliant on alternative possibilities. He argued that the freedom required for moral responsibility could be captured without reference to the freedom to do otherwise by focusing solely on the hierarchical structure of a person's desires. In this chapter, I outline and motivate Frankfurt's argument, offering a new interpretation of his positive view of the freedom required for moral responsibility. In chapter five, I criticize Frankfurt's argument claiming that it is subject to regress difficulties that undermine its cogency, and so is no

threat to PAP. In the final chapters of the dissertation, I outline, evaluate, and defend arguments that there are principled reasons to think that *any* account of the freedom required for moral responsibility not reliant on alternative possibilities will fail to capture the freedom required for morally responsible agency.

4.1 Three freedoms.

During the course of his paper, Frankfurt (1971) distinguishes between three different kinds of freedom: (i) freedom of action, or one's action being free, (ii) freedom of the will, or one's will being free, and (iii) acting freely and of one's own free will. In this section, I outline how Frankfurt describes these different freedoms and how he views their relationship to the issue of a person's moral responsibility.

Frankfurt initially distinguishes between a person's first-order desires (whose objects are actions), second-order desires (who objects are first-order desires), and his second-order volitions (which are desires that particular first-order desires should move him to act). With respect to the distinction between first-order and second-order desires, an individual's *first-order desire* is a desire to act – a desire to watch television, eat chocolate, or sleep, for instance. By contrast, a person's *second-order desire* is a desire for a desire – a desire to want to study more, or a desire to desire to exercise, for example.

A *second-order volition* is a type of second-order desire. It is a desire for a desire to move one to action. For instance, suppose a person is conflicted at the most immediate level about what he wants to do. He finds himself desiring to study but also wanting to watch television. However, he remembers that he has an exam the following day that helps him resolve his dilemma about what to do. Being of a studious disposition, and

valuing high-grades over the short-term pleasure of relaxation, he forms a desire that his desire to study, rather than his desire to watch television, should move him to act.

To further illustrate the difference between a second-order desire and a second-order volition, Frankfurt gives the example of a doctor who, in wanting to understand the symptoms of his patients but not wanting to become addicted, desires to have the first-order desire for the drug so he can experience the cravings, but does not desire to be moved by that desire because he does not actually want to take the drug (p. 9).

I want to make two other points before I outline the three kinds of freedom Frankfurt describes. First, Frankfurt calls a person's motivating first-order desire – the desire that moves him to act when he acts, or the desire that would move him to act were he to act – his *will* or his *effective desire*. Second, Frankfurt contrasts a *person* with, what he calls, a *wanton*, claiming that it is the possession of second-order desires – and second-order volitions, in particular – that is "essential" (p. 10) to personhood.[15] By contrast, a wanton is a being who "does not care about his will" (p. 11), for although he has first-order desires, he has no preferences about which of these desires he would like to move him to act. Because he lacks second-order volitions, a wanton is not a person. (Frankfurt allows that a wanton may have second-order desires that are not volitions.)

When it comes to his account of freedom, Frankfurt's foils are the classical compatibilist accounts of freedom and responsibility popular during the first half of the twentieth century (e.g., Ayer, 1954; Moore, 1912). According to these views, freedom is

---

15 At one point in his essay, Frankfurt says that it is an individual's "having" (p. 10) second-order volitions that is essential to his being a person, whereas in earlier parts of the paper he claims that it is merely the "capacity" (p. 7), or the individual's being "able"

a matter of doing what one wants to do and being free to do otherwise in the sense that one would have done otherwise, if one had wanted to. Frankfurt argues that the problem with these accounts is that while they might capture the idea of a person having freedom of action, they "miss … entirely … the peculiar content of the quite different idea of an agent whose *will* is free" (p. 14). With this as background, Frankfurt distinguishes freedom of action from freedom of the will as follows:

> … freedom of action is (roughly, at least) the freedom to do what one wants to do. Analogously, then, the statement that a person enjoys freedom of the will means (also roughly) that he is free to want what he wants to want. More precisely, it means that he is free to will what he wants to will, or to have the will he wants. Just as the question about the freedom of an agent's action has to do with whether it is the action he wants to perform, so the question about the freedom of his will has to do with whether it is the will he wants to have (p. 15).

These remarks raise the question of what Frankfurt means by speaking of an individual as being *free* to so-and-so, in these senses. After all, with respect to freedom of action, "there is a sense," as John Martin Fischer (1986) points out, "of 'having the power to do what one wants' on which one would have this power insofar as one had the power to do what one *actually* wants (but nothing else)" (p. 44, footnote 28). Similarly, with respect to freedom of the will, there is a sense in which if an individual wills what he wants to will then he must have been *free* to do so, even if he could not have willed otherwise.

(p. 6), to form second-order volitions that is required for personhood. I set this ambiguity aside.

Does Frankfurt mean by freedom of action and freedom of the will only a 'one-way' power that does not rely on alternative possibilities or considerations as to what the individual *could* have done? I do not think so. Later in his essay, Frankfurt claims that for an individual's will to be free – or for him to have freedom of the will, a term Frankfurt treats as synonymous – it must be true that "with regard to any of his first-order desires, he is free either to make that desire his will or to make some other desire his will instead. Whatever his will, then, the will of a person who is free could have done otherwise; he could have done otherwise than to constitute his will as he did" (pp. 18-19). Freedom of the will, then, requires in some sense the freedom to *will* otherwise. And, since Frankfurt claims that freedom of the will and freedom of action should be treated analogously, I take it that he thinks of a correct analysis of freedom of action as involving a freedom to do otherwise – in this case, a freedom to *act* otherwise.

Frankfurt argues that an individual fails to have freedom of action when he is "not free to translate his desires into actions or act according to the determinations of his will" (p. 15). Presumably he has in mind situations in which a person is motivated to do something – to take an elevator, for instance – yet is prevented from acting, from 'translating his will into action,' by, say, being physically restrained by another person. We can identify the individual's will or motivating desire in this case because it is his desire to take the elevator that would move him to act, were he able to act (recall Frankfurt's definition of a person's will as not merely the first-order desire that moves him to act, but also the desire that would move him to act when or if he acts).

Finally, on the difference between freedom of action and freedom of the will, Frankfurt argues that freedom of action is neither a necessary nor sufficient condition for

freedom of the will (pp. 14-15). A person can have freedom of the will without possessing freedom of action because he might be free to have the motivating desires he wants to have while being unable to act from that motivating desire. As Frankfurt would put it, he might be unable to translate his will, the desire that motivates him, into action. A person locked up in chains would be a good example of someone who lacks freedom of action but may well have freedom of the will. After all, while the chains prevent the prisoner in a straightforward sense from being free to act as he desires, he would presumably be free to will whatever he wants to will. His problem is that, being locked up, he is not free to translate his will into actual behavior.

As for freedom of action not being sufficient for freedom of the will, I give two examples. First, plausibly, a wanton has freedom of action but not freedom of the will because while it may be true that he is free to act as he desires, the fact that he does not have second-order volitions means that he is not free to be motivated by desires that he wants to be motivated by. The wanton lacks freedom of the will "by default" (p. 15), as Frankfurt puts it. Second, Frankfurt gives an example of an unwilling drug addict, an addict who does not want to be an addict and struggles (unsuccessfully) against his addiction. The addict forms a second-order volition towards his desire *not* to take the drug, but it is his addictive desire *to* take the drug that moves him to action. I think that the unwilling addict has freedom of action but not freedom of the will. He possesses freedom of action because he is free to act as he desires and, plausibly, he is free to do otherwise in the sense that had some other desire motivated him, he would have acted from that desire. But, he lacks freedom of the will because the irresistibility of his desire for the drug ensures that, in a straightforward sense, he is not free to will what he wants

to will. The unwilling addict wants his desire *not* to take the drug to be the desire that motivates him but he is unable to form his will accordingly – he is unable to 'make' this desire his will, so to speak – because of the strength of his desire *for* the drug. The difference between the person locked in chains (who possesses freedom of the will) and the unwilling addict (who does not) is that the prisoner is free to have the will he wants to have in a way in which the unwilling addict is not.

Having discussed the distinction Frankfurt makes between freedom of action and freedom of the will, I turn to the third kind of freedom Frankfurt describes, the freedom to act freely and of one's own free will. Unlike the other two freedoms, this freedom does not require that the person could have done otherwise. Frankfurt introduces this third freedom by reference to the previous two:

> It is a mistake, however, to believe that someone acts freely only when he is free to do whatever he wants or that he acts of his own free will only if his will is free. Suppose that a person has done what he wanted to do, that he did it because he wanted to do it, and that the will by which he was moved when he did it was his will because it was the will he wanted. Then he did it freely and of his own free will (p. 19).

So, for Frankfurt, a person *acts freely and of his own free will* if and only if the first-order desire from which he acts is his will because it was the will he wanted.[16] Frankfurt then argues that this kind of freedom is all the freedom required for a person to be morally responsible for his behavior. Moral responsibility requires neither freedom of

---

16 Strictly speaking, the quotation only specifies that meeting these conditions is sufficient for acting freely and of one's own free will. I set this complication aside.

action (and so not the freedom to act otherwise) nor freedom of the will (and so not the freedom to will otherwise). All it requires is that an individual acts from a desire that is his will because it was the will he wanted. Describing a person who acts with this kind of freedom, Frankfurt writes:

> Even supposing that he could have done otherwise, he would not have done otherwise; and even supposing that he could have had a different will, he would not have wanted his will to differ from what it was. Moreover, since the will that moved him when he acted was his will because he wanted it to be, he cannot claim that his will was forced upon him or that he was a passive bystander to its constitution. Under these conditions, it is quite irrelevant to the evaluation of his moral responsibility to inquire whether the alternatives that he opted against were actually available to him (p. 19).

This condition on the freedom needed for moral responsibility comprises Frankfurt's *indirect* attack on PAP. In his earlier paper, Frankfurt (1969) argued directly that PAP is false by way of an apparent counterexample known as a Frankfurt case. His route to the claim that PAP is false is different here. In this later paper, Frankfurt is attempting to undermine the truth of PAP indirectly by describing a kind of freedom that does not rely on alternative possibilities but is sufficient, in his view, to capture the freedom required for moral responsibility. If it is true that a person acts with the freedom required for moral responsibility if he is motivated by a desire that is his will because it was the will he wanted, then PAP would be false. This is because a person can act with this kind of freedom without needing to be able to do otherwise in any sense.

4.2 Frankfurt as a compatibilist.

Frankfurt is usually identified as a compatibilist about free will and moral responsibility on the one hand and the truth of causal determinism on the other. In his first remarks about this in his 1971 paper, Frankfurt writes:

> My conception of freedom of the will appears to be neutral with regard to the problem of determinism. It seems conceivable that it should be causally determined that a person is free to want what he wants to want. If this is conceivable, then it might be causally determined that a person enjoys a free will (p. 20).

What is the relationship concerning the compatibility between each of Frankfurt's three conceptions of freedom – freedom of action, freedom of the will, and acting freely and of one's own free will – and the truth of determinism? I see how on Frankfurt's view it could be causally determined that a person 'acts freely and of his own free will' and so causally determined that a person acts with the freedom required for moral responsibility. This is because I see how it could be causally determined that the desire that motivates a person was his will because it was the will he wanted. So far as his account of the freedom required for moral responsibility is concerned, Frankfurt is a compatibilist.

However, the relationship between his other two freedoms – freedom of action and freedom of the will – and the truth of determinism is more complicated. In speaking of freedom of action and freedom of the will as simply the 'freedom to do what one wants to do' and the 'freedom to will what one wants to will,' Frankfurt leaves it open that these freedoms *could* be compatible with the truth of determinism. This is because the freedoms to do otherwise that are required for freedom of action and freedom of the

will – the freedom to act otherwise and will otherwise, respectively – could, presumably, be given compatibilist-friendly interpretations. However, if Frankfurt wants to claim that these two freedoms really *are* compatible with the truth of determinism, then he has more work to do. He would need to provide and defend analyses of the freedoms to act and will otherwise that are compatible with the truth of determinism. In his only remarks about this, Frankfurt writes, "it is a vexed question just how 'could have done otherwise' is to be understood in contexts such as this one" (p. 19). If Frankfurt believes that this is a vexed question and in the absence of providing a compatibilist-friendly account of the freedom to do otherwise, then he would not be entitled to claim that freedom of action and freedom of the will are in fact compatible with the truth of causal determinism.

It is clear that Frankfurt is a compatibilist about the freedom required for moral responsibility, a kind of freedom that, on his view, does not involve the freedom to do otherwise. Should we think that Frankfurt would *also* want to argue that his conceptions of freedom of action and freedom of the will are compatible with the truth of causal determinism, then we must interpret Frankfurt as needing to claim that the alternative possibilities required for such freedom are compatible with determinism. He would then be, what John Martin Fischer (2006: 83) calls, a "compatibilist semicompatibilist". A semicompatibilist believes that "moral responsibility is compatible with causal determinism, even if causal determinism is incompatible with freedom to do otherwise" (Fischer 2006: 78). A *compatibilist semicompatibilist* interprets the freedom to do otherwise in a compatibilist-friendly manner, even though he does not think that such freedom is required for moral responsibility. An *incompatibilist semicompatibilist*, by

contrast, thinks that the freedom to do otherwise, while not necessary for moral responsibility, is incompatible with the truth of causal determinism.[17]

4.3 Moral responsibility and second-order volitions.

Let us turn back to the account of freedom Frankfurt believes is required for a person to be morally responsible for what he does. In Frankfurt's view, a person acts with the freedom required for moral responsibility if and only if the desire "by which he was moved when he did it was his will because it was the will he wanted" (p. 19). The meaning of this condition is not as clear as it might first appear. What is in need of clarification is the precise nature of the relationship Frankfurt envisages between the second-order volition of a morally responsible agent and his motivating desire. To see this, we can ask: what does it mean to say of a person who acts with this kind of freedom that the desire that moved him 'was his will because it was the will he wanted'?

On the most natural interpretation of what Frankfurt intended to communicate, to say that the desire that moved the person 'was his will because it was the will he wanted' is to say that his desiring to be moved by that desire *caused* that desire to move him. On this interpretation of Frankfurt's condition, the point of speaking of second-order volitions in connection with a person's moral responsibility is to ensure that a person has some sort of causal control over the first-order desires that motivate him. He exercises control over his own motivations, on this picture, in the sense that it is his desiring to be moved by a particular desire that causes him to act from that desire.

---

17 Fischer says, "I am an agnostic semicompatibilist, although I am perhaps a latently incompatibilist semicompatibilist" (2006: 83).

While on the face of it, this causal reading is the most natural interpretation of what Frankfurt had in mind in describing the conditions for moral responsibility – and a good number of philosophers interpret him in this way[18] – this reading omits a key feature of second-order volitions that Frankfurt stresses in his article: their place in grounding or explaining the sense in which a person can 'identify' himself with, and thereby more truly own, his first-order desires.

Frankfurt makes these points about a person's identification with his desires in connection with his example of the unwilling drug addict. This addict is a person who takes the drug in spite of himself. He has both a desire *to* take the drug and a desire *not* to take it. While both these desires "are his, to be sure," the addict "identifies himself … through the formation of a second-order volition with one rather than with the other of his conflicting first-order desires," making "one of them more truly his own and, in doing so," withdrawing "himself from the other" (p. 13). It is in virtue of his identification with his desire not to take the drug that we can speak of the addict as being 'unwilling,' since the desire from which he acts – his desire for the drug – is not the desire that he wants to be moved by.

In these remarks, Frankfurt seems to be distinguishing between two different senses in which a person's desires can be said to be 'his'. On the one hand, there is the brute or trivial sense in which all of a person's desires are his desires simply in virtue of

---

18 Eleonore Stump (1988), for instance, writes in this connection that Frankfurt requires that a morally responsible agent "has the first-order volitions he has *because* of his second-order volitions (that is, his second-order volitions have, directly or indirectly, produced his first-order volitions)" (p. 397). Fischer (1986) and Kane (1996) also describe Frankfurt's view in terms of an individual's morally responsible behavior as being caused by his second-order volitions in this way.

their being a part of his mental life. In this respect, the unwilling addict can be said to 'own' both of his conflicting first-order desires: his desire *to* take the drug, and his desire *not* to take it. On the other hand, there is a stronger – what we might call, a more genuine, true, or complete – sense of ownership a person can have with respect to his desires. It is when a person *identifies himself* with a desire, something he does by desiring to be moved by that desire, that the desire belongs to him in this stronger sense.

Given the important role that second-order volitions have for Frankfurt in explaining how a person can identify himself with his desires, this suggests an alternative reading of Frankfurt's condition for morally responsible agency. On this alternative picture of Frankfurt's condition, to say that the desire that moved the person 'was his will because it was the will he wanted' is not to say that his desiring to be moved by that desire caused him to act from it. Rather, it is to say that it is by desiring to be moved by that desire that the person identifies himself with his motivating desire, and so when he acts from it, he acts from a desire that is truly 'his'. We can bring out this alternative reading by supplying the following emphasis to Frankfurt's remark: a person acts with the freedom required for moral responsibility if and only if "the will by which he was moved when he did it was *his* will because it was the will he wanted" (p. 19). On this ownership interpretation of Frankfurt's condition, the point of speaking of second-order volitions in connection with a person's moral responsibility is not to ensure that the person has some sort of causal control over the motivations of his behavior but, instead, to secure the idea that a morally responsible agent acts from desires he genuinely owns, motivations that are truly 'his'.

It is unclear which interpretation of his condition Frankfurt intended. His most pertinent remarks about it are contained in his description of the *willing* drug addict, an addict who, unlike the unwilling addict, is "altogether delighted with his condition" (p. 19). He is a 'willing' addict in the sense that it his desire *to* take the drug, rather than his desire *not* to take it, that he wants to be moved by. Frankfurt also says that this addict meets his condition for the freedom required for moral responsibility since his desire for the drug is his will because it was the will he wanted. On the one hand, Frankfurt describes the willing addict's situation "as involving the overdetermination of his first-order desire to take the drug. The desire is his effective desire because he is physiologically addicted. But it is his effective desire also because he wanted it to be" (pp. 19-20). This overdetermination remark implies that Frankfurt has the causal reading in mind. It seems that he wants to convey the thought that there are two things true of the addict that are individually sufficient to ensure that his desire for the drug is effective: that his desire is addictive, and that his desire is the one he wants to be moved by.[19]

However, Frankfurt follows this remark by suggesting that "by his second-order desire that his desire for the drug should be effective, he has made his will his own" (p. 20). This suggests the possibility of the ownership reading, that the point of citing second-order volitions in connection with a person's moral responsibility is to secure the

---

19 Fitting the overdetermination remark with the ownership reading is more difficult. If what Frankfurt had in mind is that it was overdetermined that the willing addict identifies himself with his desire to take the drug, then this means that Frankfurt would be holding that a sufficient condition for a person to identify himself with a desire is that that desire is addictive. But, in the case of the *unwilling* addict, whose desire for the drug is addictive, Frankfurt claims that he does *not* identify himself with this desire.

sense in which, when acting with the freedom required for moral responsibility, a person acts from desires that are truly his.

4.4 Second-order volitions and self-expression.

What is most interesting about these opposing interpretations of Frankfurt's condition is not simply that there is an ambiguity in his account that Frankfurt does not clearly resolve, but that the conflict brings to light the broader issue of the general picture of freedom Frankfurt wants to capture by casting a person's moral responsibility in terms of his second-order volitions. We know that Frankfurt rejects the claim that moral responsibility requires the freedom to do otherwise. But what kind of freedom, in the broadest sense, *does* Frankfurt believe moral responsibility requires?

Though Frankfurt is not explicit in answering this question, I follow many philosophers in viewing him as arguing that the freedom people must act with in order to be morally responsible for their behavior is, what we can call, the *freedom of self-expression*. On the self-expression model, people act with the freedom required for moral responsibility to the extent that their actions reflect who they really *are*, as people – to the extent that their actions reflect their 'true,' 'deep,' or 'real' selves, as it is sometimes put.[20] Interpreting Frankfurt's view along self-expression lines implies that the point of speaking of a person's second-order volitions in connection with his moral responsibility is to secure the sense in which the desires from which he acts are, in some

---

20 Susan Wolf (1987, 1990) coined the terms 'deep' and 'real' self in this connection. Others who interpret Frankfurt in a similar way include Nomy Arpaly and Timothy Schroeder (1999), Laura Waddell Ekstrom (2005), Robert Kane (2005), Elinor Mason (2005), and Kasper Lippert-Rasmussen (2003).

way, a fundamental part of his self; in acting from these desires, the person's action will then reflect who he really is, as a person.

This understanding of Frankfurt's project can help us decide whether the causal reading or the ownership reading is the best way to interpret Frankfurt's condition. In particular, we can ask what makes it the case that the desire that motivates the morally responsible agent is part of his fundamental self? Is it the fact that he causes himself to be moved by it (the causal reading) or the fact that he 'identifies' himself with it (the ownership reading)? On the face of it, I think that the ownership reading gives the more plausible answer to this question. The difference can be put in terms of whether we should think of a person's fundamental self as composed of those desires with which he identifies himself or those desires that he causes himself to act from. The problem with the latter account, the causal reading, is that there are surely desires that are part of a person's fundamental self that are *not* desires that, for one reason or another, lead the person to act.

As an example, consider Frankfurt's case of the unwilling drug addict. Frankfurt implies that it is this addict's desire *not* to take the drug, rather than his desire *to* take it (which is the desire that moves him), that is the more fundamental part of him. According to the explanation implied by the causal reading, it is when a desire is one from which a person *acts* – one that he causes himself to act from, more specifically – that a desire becomes part of the person's 'true' self. This provides no way to explain how it is the unwilling addict's desire *not* to take the drug, which is a desire that *fails* to motivate him, that is part of his fundamental self. Perhaps an alternative explanation for these kinds of cases can be given, but all things equal, it is preferable to have a consistent

explanation across all cases. The account that falls out of the ownership reading – that what makes it the case that a desire is part of a person's fundamental self is that he identifies himself with it – covers both kinds of case. It alone explains how a desire can be part of a person's 'true' self whether or not it is one from which the person acts.

Reflection on the unwilling addict case brings up a broader problem that threatens to undermine *any* account of moral responsibility based on the idea of self-expression. This is the issue of how, pre-theoretically, we should identify those parts of a person's self that are more fundamental to who he is than other parts. After all, there is a straightforward sense in which who this person 'really is' is a drug addict, even if he identifies himself with his desire not to take the drug rather than with his desire to take it. 'He really is a drug addict,' we might say, 'even though he doesn't want to be one.' To develop this point, we might distinguish between 'who a person really is' on the one hand and 'who a person really wants to be' on the other, and insist – against the self-expression theorist – that it is not at all obvious that we should understand the former in terms of the latter, that is, understand who a person really is in terms of who he really wants to be.

I think that Frankfurt and his defenders should reply by granting that while it is true that the unwilling addict *is* an addict no matter what the content of his second-order volitions, there is also a *further* understanding of the idea of a person's 'true' or fundamental self in which his desire for the drug would not be a part. This is the idea of a person's fundamental self that we recognize when we say, in excusing the addict, 'it wasn't really him, it was the drug that was moving him.' (Arpaly and Schroeder [1999] cite this excuse in this connection.) The idea here is that there is some part of this person's self, besides his addiction, that is more fundamental to who he really is. And it

is *this* part of the person's self, the self-expression theorist should insist, that a person's action must reflect in order for him to be morally responsible for it.

4.5 The Willing drug addict.

I want to finish this chapter by looking in more detail at Frankfurt's example of the willing drug addict which throws further light on Frankfurt's theory. Frankfurt explicitly introduces this example to illustrate his claim that "it is not true that a person is morally responsible for what he has done only if his will was free when he did it" (p. 18). Though the willing addict could not have willed otherwise, his addiction ensuring that his desire for the drug is his will, his relationship to his will – its being 'his will because it was the will he wanted' – means that, in Frankfurt's view, he acts with the freedom required for moral responsibility. What is striking about this case is that it seems just as much an apparent counterexample to PAP as the 'Frankfurt cases,' the examples Frankfurt developed in his earlier paper, 'Alternate possibilities and moral responsibility (1969). In both kinds of case, a person apparently acts with the freedom required for moral responsibility despite the fact that some feature of the situation – Black's presence or the addict's addiction – allegedly ensures that the person could not have done otherwise.

What is also striking about the cases is not simply the symmetry between them but also the difference in the extent to which these examples have been deployed as counterexamples to PAP. While it is commonplace to use the Frankfurt cases as a counterexample to this principle, one rarely – if ever – finds the willing addict case being

used to draw the same conclusion.[21] Why have PAP rejecters neglected to use the willing addict case as a counterexample to PAP?

I suspect that those wanting to reject PAP have failed to use the willing addict case to undermine the principle because they have the intuition that his addiction takes away some of his moral responsibility even though they do not have the corresponding intuition that Black's presence in the Frankfurt case takes away some of Jones' responsibility. Indeed, some philosophers suggest that an addict cannot be morally responsible for taking a drug under any circumstances, no matter what the nature of his higher-order satisfaction with his behavior. Don Locke (1975), for instance, argues that the distinction between the moral responsibilities of the unwilling and willing addicts is "implausible given that they are both equally addicts." He goes on to say that "while I can see a respect on which the willing addict might be regarded as more reprehensible, in that his desires are as depraved as his actions, this does not seem to make any difference to his responsibility for his behavior" (p. 100).[22] Of course, one question that should be

---

21 For instance, none of the essays in a recent book collection devoted to Frankfurt's work concerning the relationship between moral responsibility and alternative possibilities (Widerker & McKenna, 2003) even mentions the willing addict case.

22 I also find this uncertainty in intuition about whether or not an addiction takes away *in principle* an addict's moral responsibility when taking the drug in Fischer's work. In early work, after discussing the contrast between the unwilling and the willing addict, Fischer (1986) claims that "it would be desirable to have a theory of responsibility that would explain why such agents as … the 'happy addict' are responsible for their actions, although they couldn't have done otherwise" (p. 43). In more recent work, however, he seems to back away from this desideratum, at least implicitly. For instance, writing with Mark Ravizza (Fischer & Ravizza, 1998: 48), he speaks of drug addicts as being those who we intuitively think of as *not* responsible for their actions. They consider a drug addict who takes a drug and argue that the mechanism that led to his action cannot be described as any type of normal 'deliberation' because of it were, then the addict would,

pressed against philosophers who believe that PAP is false but also think that the addict's addiction takes away some of his responsibility is what is it *about* his addiction that attenuates his moral responsibility? The answer on their part had better not be that the addiction takes away from the person's responsibility because it deprives him of alternative possibilities. But if it is not this feature of his addiction that makes it morally relevant, then what other explanation could there be?

Setting this question aside, I want to finish this chapter by developing an argument against PAP using the willing addict case, one that rejects the assumption made by Locke that an addict can never be responsible for his drug-taking under any circumstances. The argument has two stages.

First, the PAP rejecter should use the willing drug addict to challenge the claim that a drug addict can never be morally responsible for taking a drug. Comparison of the willing addict case with the *unwilling* addict case should lead us to consider the question, 'Under what conditions should an addiction exculpate?' Comparing the willing addict case with the unwilling addict case might elicit the intuition that an addiction will not *always* exculpate. The willing addict wants his desire for the drug to move him to act. By desiring that it should move him, the addict has identified himself with this desire. This act of identification means that he will be frustrated should any other desire besides his desire for the drug move him to act. By contrast, the *unwilling* addict acts from a desire that he does not want to be moved by. In fact, he wants a different desire to move him, his desire not to take the drug. To this extent, the unwilling addict is frustrated with

---

according to their view, be morally responsible for his action (and they do not want this result).

the nature of his will. The difference in these two cases, particularly with respect to the satisfaction these addicts have with their motivating desires, might lead us to reasonably cast doubt on the claim that addiction *always* exculpates. Having conceded that there might be circumstances in which an addiction may not excuse the person, let us move to the second step of the argument.

The second step of the argument consists in arguing that there is a *morally relevant* difference in the responsibility both addicts bear towards their actions. The unwilling addict acts from a desire that he does not identify himself with. This ensures, as Frankfurt puts it, that "the force moving him to take the drug is a force other than his own" (p. 13). It might seem for this reason – the reason that the addict is not being moved by a desire that is truly his – inappropriate to blame him for taking the drug. Yet the willing addict *does* identify himself with the source of his action. He wants to be moved by his desire for the drug. So, the argument goes, it could be appropriate to blame the willing addict for this reason, that his motivating desire is a desire that is a true part of him, one he has genuine ownership over.

This difference in the extent to which these two addicts are moved by desires that are fully theirs is, by this line of reasoning, a morally relevant difference. If this is right, then we have a counterexample to PAP, since neither addict has the freedom to do otherwise. By highlighting the difference in degree of ownership the two addicts have with respect to their wills, we have constructed an argument in which, it is alleged, a person can act with the freedom required for moral responsibility despite lacking the freedom to do otherwise.

How persuasive is this argument? Should it lead us to abandon PAP? I suspect not. For many, myself included, the intuitions supporting PAP brought to light by asking the question, 'if you want to blame the willing addict, what should he have done instead?' would count against thinking the addict morally responsible for his drug-taking. In fact, reflection on the fact that there is no good answer to what the addict should have done instead should lead us, I think, to embrace Locke's remarks about the responsibility of an addict. Because they could not have done otherwise, neither addict can be morally responsible for what they did. Yet the difference in the relationship both addicts bear towards their wills might make a morally relevant difference in the assessment of their moral *character*. The willing addict can be judged more reprehensible, as Locke suggests, if not morally responsible. (I explore the difference between judgments of moral responsibility on the one hand and judgments of moral character on the other in chapter six.)

Chapter Five: Evaluating Frankfurt's Hierarchical Account

In the previous chapter, I offered what I believe to be the most plausible interpretation of Frankfurt's positive account of the freedom required for moral responsibility. In this chapter, I want to evaluate it. I focus on two criticisms. First, I assess whether a person must identify himself with his motivating desire in order to be morally responsible as Frankfurt suggests. I focus in particular on apparent counterexamples that suggest that identification is not necessary, though I argue that Frankfurt and his defenders have the apparatus to avoid these alleged counterexamples.

Assuming that identification is necessary for responsibility, the second criticism I look at focuses on whether Frankfurt's condition that the first-order desire from which a person acts is his will because it was the will he wanted is sufficient to capture the way in which a morally responsible agent identifies himself with his motivating desire. I argue that this criticism is much more troubling for Frankfurt's account than the first one. Criticisms that Frankfurt's condition is not sufficient for identification are not rare, but I try to break new ground by arguing that his account is not subject to one regress difficulty, as is commonly argued, but to two independent regress problems.

5.1 Is Frankfurt's condition necessary for moral responsibility?

The first criticism to be considered concerns whether identification is necessary for moral responsibility. R. Jay Wallace (1994) claims that Frankfurt's condition is not necessary for the freedom moral responsibility requires on the grounds that we blame people even when they "act spontaneously" or "against their better judgment" even

though “in none of these cases do their actions reflect their higher-order identifications” (p. 264). Recasting this a little, we can ask whether a person must identify himself with motivating desire in order to be blameworthy or praiseworthy for what he does.

Of the cases mentioned by Wallace, the claim that weak-willed actions – that is, behaving against one’s better judgment – are counterexamples is unconvincing. When a person knowingly acts against his moral judgment, it does not follow – as Wallace and other (for instance, Kane [2005]) have claimed – that the individual will be failing to act from a desire that he has identified with, a desire that is his will because it was the will he wanted. This is because Frankfurt distinguishes between a person’s identification on the one hand, and his moral judgment on the other. With this difference in hand, he denies that “a person’s second-order volitions necessarily manifest a *moral* stance on his part towards his first-order desires.” In fact, “it may not be from the point of view of morality that the person evaluates his first-order desires” (p. 13, footnote 6). On Frankfurt’s view, a person may identify himself with a certain desire knowing full well that it would be morally wrong for him to act from it. Perhaps he does not care about morality, about behaving morally, or perhaps he accepts his immoral inclinations. Either way, Frankfurt can hold a person responsible for acting against his moral judgment without undermining his own theory so long as the person identifies with his motivations.

Wallace’s suggestion that acting spontaneously – or acting ‘on a whim,’ as Kasper Lippert-Rasmussen (2003) puts it in his argument against Frankfurt – is a counterexample to Frankfurt’s condition is a more promising criticism. If a person acts spontaneously, without having identified himself with his motivating desire, it does seem, on the face of it, “implausible” (p. 372), as Lippert-Rasmussen suggests, to deny that the

person is responsible simply because he fails to form a second-order volition. One line of response here is to point out that Frankfurt insists that a person can form his second-order volitions in a "capricious and irresponsible" manner, so identify himself by "giving no serious consideration to what is at stake" (p. 13, footnote 6). With such little restriction on the conditions of their formation, perhaps Frankfurt and his defenders could insist that in these instances of apparently spontaneous behavior, the person really *does* identify himself with his motivating desire, albeit spontaneously and without a period of sustained reflection.

But this reply is no good in the face of a critic who wants to stipulate an example in which a person acts from a first-order desire without having formed a second-order volition of any kind, neither towards his motivating desire nor towards some other desire. I think that Frankfurt's best response here is to fight intuition with intuition. On the one hand, there is no doubt some intuitive pull to think that such a person ought to be worthy of blame or praise for his behavior. On the other hand, though, Frankfurt and his defenders should insist that there are other intuitions supporting the claim that the proper judgment is that a person who has not identified himself with his motivations is *not* morally responsible. The point of speaking of identification, remember, is to secure the sense in which morally responsible behavior reflects the self. With this as background, consider the 'wanton,' an individual who lacks second-order volitions, and so has no genuine self to which his behavior can be attributed, no proper self that his actions can reflect. The wanton's behavior is animalistic, simply a result of his brute urges. When a *person* acts from mere desire alone without identifying himself with his motivations, then while he *has* at least the possibility of a genuine self that his actions could reflect (unlike

the wanton), on this occasion his behavior is importantly 'wanton-like'; it is simple brute behavior that does not express the self. And since we have the intuition that the wanton is not responsible for what he does on the grounds that his behavior is not self-expressive, this gives some reason to think that a *person* is not morally responsible for his spontaneous actions for similar reasons, when his actions are not expressive of his self.

5.2 Can Frankfurt's condition capture identification?

In the previous section, I argued that Frankfurt's suggestion that identification is necessary for moral responsibility can be defended against apparent counterexamples. Granting that some kind of identification is necessary, I now want to consider whether or not Frankfurt's condition is *sufficient* to capture the sense in which a person identifies himself with his motivating desire when morally responsible for what he does.

Some argue that Frankfurt's condition is not sufficient on the grounds that second-order volitions are, after all, 'merely' desires; "since second-order volitions are themselves simply desires," writes Gary Watson (1975), "to add them to the context of conflict is just to increase the number of contenders; it is not to give any special place to any of those in contention" (p. 218). I am not entirely clear about how this criticism should be understood. On one reading of it, what is objectionable is the fact that second-order volitions share with their first-order objects the property of being desires, and a desire no matter what its level in a hierarchy lacks the necessary significance required to confer identification. (Watson might have this argument in mind since he claims that it is only a person's *values*, as opposed to his desires, that carry enough significance to speak for a person in this respect.) If this is how the criticism is meant to be taken, then I do not

find it persuasive. It is especially unclear why we should think that the fact that second-order volitions are kinds of desire would prevent them from explaining the sense in which some of a person's first-order desires are more truly his own than others.

Another way to read Watson's remarks, though, is as drawing attention to the fact that speaking of second-order volitions – desires for desires – opens up the possibility of desires of a higher order than the second level. This phenomenon may cast doubt on the sufficiency of Frankfurt's original condition, as I now want to explain. In fact, I think that Frankfurt is trying to express this kind of worry in the following passage from his original paper:

> [One] complexity is that a person may have, especially if his second-order volitions are in conflict, desires and volitions of a higher-order than the second. There is no theoretical limit to the length of the series of higher and higher orders; nothing except common sense and, perhaps, a saving fatigue prevents an individual from obsessively refusing to identify with any of his desires until he forms a desire of the next higher order. The tendency to generate such a series of acts of forming desires, which would be a case of humanization run wild … leads towards the destruction of the person (p. 16).

This passage is suggestive but opaque. Particularly puzzling is Frankfurt's claim that a person can 'refuse to identify himself with a desire until he forms a desire of the next higher order.' After all, according to Frankfurt's account, to identify oneself with a desire just *is* to form a higher-order volition towards it. However, the worry I think that Frankfurt is trying to express here is that once we use the idea of a hierarchy of desires to explain identification, we open up the possibility that a person could reflect on the

desirability of his second-order volition from the third-order level thus undermining the efficacy of that volition to explain the way in which his motivating desire is truly his. I think that Frankfurt is aware that, as it stands, his account of identification relying on second-order volitions alone is incomplete; identification requires something more than desiring to be moved by the particular desire.

To explain, consider an example of a hard-working student who finds himself with a stray desire to get drunk as well as his usual desire to study. As is his way, he forms a second-order volition towards his desire to work. He wants his desire to study rather than his desire to drink to move him to act. He meets Frankfurt's original condition, so according to Frankfurt's original picture, he acts with the freedom required for moral responsibility should he be moved by his desire to study. However, we can see that Frankfurt's condition is not sufficient to explain the student's identifying himself with his motivating desire. Imagine that the student is conflicted about his moral character, about the kind of person he is, and being fed up with his strict work ethos, he desires, from the third-order level, *not* to desire to be moved by his desire to study. He desires to rid himself of his second-order volition, one that reflects, as it were, his now unwanted studious disposition. While on the one hand, he wants his desire to study to move him, on the other hand, this is a second-order volition that because of his personal conflict he wishes not to have. But, how can his second-order volition explain his identification with his motivating desire under these circumstances if it is a desire that he does not want to have? Surely, any authority his volition would otherwise have to explain how his motivating desire is truly his is undermined by the fact that it is volition the person desires not to have.

While the student example shows that Frankfurt's condition is not sufficient to capture the sense in which a person binds himself with his motivating desire when morally responsible for what he does, I think that the difficulty Frankfurt gestures at in the passage I quoted is broader than that I have sketched in the conflicted student example. I think that the problem does not simply turn on the fact that the student desires not to have the desire to be moved by his desire to study. Rather, the difficulty arises as soon as a person "put[s] his relationship to it [his second-order volition] in question," as Frankfurt (2002) puts it in a later paper, something he does simply by "reflecting on" it (p. 86). By mentally 'stepping back,' so to speak, and reflecting on whether or not he wants to desire to be moved by his desire to work, the student has undermined the authority his volition would otherwise have to explain his identification with his desire to study. After all, how can his second-order volition explain how this is a desire he more truly owns if it is a desire that the person is questioning whether or not he wants to have?

The upshot of these considerations is that simply desiring to be moved by a desire is not enough for a person to identify himself with his motivating desire since he could meet this condition while also questioning whether or not he *wants* to want to be moved by it. So long as this possibility is open, Frankfurt's account of identification is incomplete.

5.3 Frankfurt's response.

Following the passage of his that I quoted, Frankfurt offers what I take to be a solution to the difficulty I have been discussing, an addition to his original condition

intended to close off the possibility that a person could undermine his identification by questioning the desirability of his second-order volition. He writes:

> It is possible, however, to terminate such a series of acts [i.e., the formation of desires of a higher order than the second level] without cutting it off arbitrarily. When a person identifies himself *decisively* with one of his first-order desires, this commitment 'resounds' throughout the potentially endless array of higher orders. Consider a person who, without reservation or conflict, wants to be motivated by the desire to concentrate on his work. The fact that his second-order volition is a decisive one means that there is no room for questions concerning the pertinence of desires or volitions of higher orders. Suppose the person is asked whether he wants to want to concentrate on his work. He can properly insist that this question concerning a third-order desire does not arise. It would be a mistake to claim that, because he has not considered whether he wants the second-order volition he has formed, he is indifferent to the question of whether it is with this volition or with some other than he wants his will to accord. The decisiveness of the commitment means that he has decided that no further question about his second-order volition, at any higher order, remains to be asked (p. 16).

As I read this passage, Frankfurt suggests in the last line of his remarks that his original account should be supplemented with the condition that in order for a person to identify himself with his motivating desire (and hence act with the freedom required for moral responsibility) he must make a decision that he has no questions about the

pertinence of his second-order volition.[23] To avoid the possibility that a person could undermine the efficacy his second-order volition would otherwise have to explain how his motivating desire is truly his by reflecting on its desirability, Frankfurt requires that a person must make a decision that he has no questions about its pertinence. Does this response successfully avoid the original problem? Is Frankfurt's new account of identification sound? I consider these questions in the next section.

5.4 Evaluating Frankfurt's appeal to decision.

Though I think that Frankfurt's response is unsuccessful, I want to make this argument by showing how a commonly cited argument against Frankfurt's response is unsound. In a famous critique of Frankfurt's article, Watson (1975) argues that either Frankfurt's reply here:

> … is lame or it reveals that the notion of a higher-order volition is not the fundamental one. We wanted to know what prevents wantonness with regard to one's higher-order volitions. What gives these volitions any special relation to 'oneself'? It is unhelpful to answer that one makes a 'decisive commitment,'

---

23 Frankfurt goes on to say that "it is relatively unimportant whether we explain this [a person's making a decision] by saying that this commitment generates an endless series of confirming desires of higher orders, or by saying that the commitment is tantamount to a dissolution of the pointedness of all questions concerning higher orders of desire" (pp. 16-17). I think that this shows that what bothered Frankfurt about his initial account is not that a person might *actually* form desires of a higher order than the second level (for he allows that such desires might be 'implicitly' generated as a result of the decision a person makes about this second-order volition). Rather, what concerns him, and the reason for which he realized his initial account was insufficient, was that a person might want to be a moved by a particular desire yet at the same time question whether or not he wants to want to be moved by that desire.

where this just means that an interminable ascent to higher orders is not going to be permitted. This *is* arbitrary (p. 218).

This response, as stated, is unconvincing, and I am surprised when I frequently see it cited approvingly by other philosophers. It is unpersuasive because Watson misunderstands Frankfurt's solution to the problem. Making a decisive commitment, on Frankfurt's view, does not *just* mean – as Watson suggests – that 'an interminable ascent to higher orders is not going to be permitted'. Rather, it means that the person has *made a decision* that he has no questions about whether or not he wants to have his second-order volition. Since the possibility of a regress is eliminated as a result of the decision, its termination would not obviously be arbitrary, contrary to Watson's claim.

Perhaps, though, we can recast Watson's objection so as to attribute the right view to Frankfurt but preserve the idea that the way in which the possibility of the regress is terminated is nonetheless arbitrary. In particular, we can ask, for what *reason* does the person decide that he has no questions about his second-order volition? If he makes this decision for no reason, or only because he is too tired to think any harder about it, then we might think that the regress is arbitrary because the decision was *made* arbitrarily. Another way to put this criticism is to say that by deciding that he is satisfied with his second-order volition, that he has no questions about whether or not he wants it, the person "accord[s] the higher-order attitude effective authority (p. 90, footnote 1), as Frankfurt (2002) puts it in a later paper. The decision accords the volition authority to explain the way in which his motivating desire is truly his. But how can the decision accord his volition this kind of authority if it was made out of laziness or tiredness?

One way to diagnose the plausibility of this objection is that if a person were to decide that he is satisfied with his second-order volition only because he was too tired to think any harder, then this decision does not have enough significance to accord his volition the authority to explain his identification with his desire; the decision would be half-hearted, we might say. By way of response, I think that Frankfurt and his defenders should reject the conditional. They should insist that no matter *what* leads the person to make his decision, even if tiredness, his making it ensures that he is not moved to wonder whether or not he wants to have his second-order volition. Furthermore, this is what is key to avoid the possibility that he could stand back and reflect on its desirability. So long as he makes this decision, he will not be moved to question his second-order volition.

A better, and to my mind *decisive*, problem with Frankfurt's appeal to decision is that it undermines his central claim that a person's identification with his motivating desire must necessarily be understood in terms of a hierarchy of desires. Introducing the notion of decision in the way Frankfurt does raises the question of why we cannot capture the difference between those desires that a person more truly owns and those he does not by reference to him deciding that he has no questions about whether or not he wants to have these desires. On this picture, those first-order desires the person decides he has no questions about, no thoughts about whether or not he wants to have them, are those he more truly owns. By contrast, it is those first-order desires that he *has* questions about from the second-order level, those that he has not decided he is satisfied with, that are not desires he identifies himself with. Casting a person's ownership of his first-order desires in this light, by decision alone, bypasses any appeal to higher-order desires.

Gary Watson hints at this criticism in a footnote to his 1987 paper, 'Free action and free will,' when he writes that if, according to Frankfurt, second-order volitions are:

> … desires plus something else [i.e., plus a decision that the person has no questions about them], then the hierarchical account has not after all given us an account of identification; moreover, there is no reason to think that such identification is necessarily higher-order (pp. 149-150, footnote 7).

The point is that by granting decision a crucial role in his account, we can ask why we cannot dispense with second-order volitions altogether and simply couch those first-order desires a person more truly owns in terms of those desires he has decided that he has no questions about. This, though, undermines Frankfurt's central idea of understanding identification, and the freedom required for moral responsibility, in terms of a hierarchy of desires. We can emphasize the point against some remarks by David Zimmerman (1981), a defender of hierarchical views. Zimmerman writes:

> The idea is to come up with some stance which can be a source of identification and internality, but which will not generate a regress. Frankfurt believes that *decision* is such a stance (p. 359).

But if decision can be such a stance, then why must we resort to a hierarchy of desires to explain identification? Instead, we can use decision as the key explanatory notion and argue that a person identifies himself with a certain desire by virtue of making a decision that he has no questions about whether or not he wants to have that desire. It is

when a person has not made such a decision, when he questions whether he wants to have the motivations that he does, that he fails to identify himself with them.[24]

5.5 The second regress problem.

Though it has commonly been argued that Frankfurt's account is susceptible to problems of a regress of higher order desires, I think that the regress difficulty is more subtle than has been appreciated. In particular, I think that there are *two* regress difficulties to which Frankfurt's view is subject, and the solutions Frankfurt offers in both instances undermine his central argument that identification is necessarily a matter of higher order desires. On the one hand, according to the difficulty I have been sketching, Frankfurt's original condition is not sufficient for identification because it is consistent with the person putting his own relationship to his second-order volition in question by mentally stepping back to the third-order level and asking himself whether or not he wants to have that volition. Questioning whether or not he desires to have his second-order volition undermines the efficacy that volition would otherwise have to explain the way in which the person identifies himself with his desire. On the other hand, there is a *second* regress problem that is different from the one I have outlined so far. According to the second regress difficulty, Frankfurt's condition is not sufficient for identification not

---

[24] One response Frankfurt and his defenders might make to my criticism is that if we couch identification solely by reference to decisions without any appeal to second-order volitions, then we have omitted reference to the essential characteristic of a person: the possession of second-order volitions. The difference between a person and a 'wanton,' a creature who is moved by brute desire alone, is in his capacity to form preferences about what desires he wants to move him. However, we could still keep this criterion for personhood while limiting attributions of responsibility only to persons. There would

because it does not rule out a *possible* regress – the possibility that a person could question the desirability of his second-order volition – but, instead, because it entails an *actual* regress of volitions beyond the second level. Let me explain.

On Frankfurt's original view, a person's identification with his motivating desire is constituted by his forming a second-order volition towards that desire. However, presumably his second-order volition can only confer genuine ownership of his motivating desires if it *itself* is a desire he identifies himself with. By Frankfurt's condition, this would require positing a *third*-order volition to explain how he identifies himself with his second-order volition. But presumably the person's third-order volition can only convey true ownership of his second-order volition if it is *also* one he identifies himself with. And this requires positing a *fourth*-order volition, and so on *ad infinitum*. Hence, Frankfurt's account of identification commits him to an ever high order regress of desires, each necessary to confer the person's more true ownership of his prior desire.

These two regress difficulties are rarely pulled apart in critical discussions of Frankfurt's work. To give one recent example of this failure, Laura Waddell Ekstrom (2005) introduces what she calls 'the regress problem' as part of a footnote in the following way:

> The regress problem is generated as, when deciding what to do, one consults one's desires concerning what to do. But in order to avoid being wanton-like, and in the face of conflict among first-level desires, one ascends to the second level, asking oneself what one desires to desire to do. What has often been noticed is

---

though no longer be such a tight connection between responsibility attributions and the conditions for personhood.

> that nothing seems to prevent one's evaluative questioning of this second-level desire in turn, so that one asks what one desires to desire to do (p. 50, footnote 8).

This is a clear statement of the first regress problem I discussed, the difficulty that follows from the fact that a person could satisfy Frankfurt's original condition while also questioning whether or not he wants to have his second-order volition. Yet when Ekstrom elaborates on the regress problem in the main body of the text, she presents the criticism differently, in terms of the *second* regress I introduced. She writes:

> … the regress is especially problematic because ascension to higher and higher orders of desire is not only something that *might* occur, due to persistent self-doubt or intra-level conflict. Rather, it is something that must occur, as Frankfurt's original account of 'internality' requires it. What makes a first-level desire one's own … is that one has a positive endorsement of it, in the form of a second-level volition. But the second-level attitude can confer internality only if it is internal to the self; and applying the account of internality to this state requires a third-level endorsing state. What makes the third-order desire *one's own* is that one has a fourth-level desire for it; and so on (p. 50).

5.6 Responding to the second regress problem.

There is little evidence in his 1971 paper that Frankfurt realized that his own account of identification appears to commit him to the existence of an ever higher order series of endorsing desires in a morally responsible agent, each necessary to confer ownership of the prior desire. In fact, I am inclined to read much of his later work on identification as his attempt to avoid this regress. Before I survey some of this work,

though, I want to first look at a fairly straightforward reply that has been made in the face of this regress. Zimmerman (1981) responds to the difficulty by denying the truth of one of the premises on which the regress rests, the claim that a person's second-order volition can only explain his identification with his motivating desire if it – his second-order volition – is *itself* a desire that the person identifies himself with. If this premise is false, the regress would not arise for there would be no need to posit a *third*-order volition to explain the way in which the person identifies himself with his second-order volition. Zimmerman argues that if we want to provide a compatibilist-friendly account of the freedom for moral responsibility, then we have to accept that:

> … there is some point in the motivational hierarchy where the higher-order desire playing the crucial endorsing role is itself an unwilled, unendorsed part of the agent's motivational equipment, to be explained in terms of non-motivational causes, either genetic or environmental (p. 359).

This response is unlikely to persuade the critic who will likely see it as either a failure to take the regress seriously or as a *reductio ad absurdum* – if *this* is what is required in order to avoid the regress then so much the worse for Frankfurt's account. After all, it is mysterious how some feature of a person's mentality, his second-order volition in this case, could explain how a particular desire is truly 'his' if that segment of his mental life is not a part the person has any special connection to, if it is just an 'unendorsed' part of him, to use Zimmerman's phrase.

As I mentioned earlier, I view much of Frankfurt's later work on identification following his 1971 paper as an attempt by him to deal with this regress difficulty. This work is dense but unsystematic. In the following section, I want to give some indications

as to how I think it should be viewed. I ultimately argue that as with his response to the first regress problem, Frankfurt's later work undermines his original thesis since it commits him to the view that a person's identification of himself with desires need not be understood in terms of a hierarchy of higher-order desires.

5.7 Frankfurt's later work on identification.

In his 1975 paper, 'Three concepts of free action,' Frankfurt makes the following remarks about second-order volitions:

> As for a person's second-order volitions themselves, it is impossible for him to be a passive bystander to them. They *constitute* his activity – i.e., his being active rather than passive – and the question of whether or not he identifies himself with them cannot arise. It makes no sense to ask whether someone identifies himself with his identification of himself, unless this is intended simply as asking whether his identification is wholehearted and complete (p. 121).

What should we make of these claims? Frankfurt's last sentence is puzzling as even Zimmerman, one of his staunchest defenders, admits. Zimmerman says that:

> … it makes perfect sense to ask 'whether someone identifies himself with his identification of himself,' for despite the odd sound of the question, all it really comes to on … [Frankfurt's] notion of identification is the question of whether someone's second-order volition is in turn endorsed by a third-order volition, and this question is clearly in order (p. 358).

As for the idea that the question of whether or not a person identifies himself with his second-order volitions cannot arise because these volitions 'constitute' his being

active rather than passive, I am ambivalent. Frankfurt's idea seems to be that this question cannot arise since a person necessarily identifies himself with them by virtue of these volitions constituting his activity. But the crucial question of what it means for a person's second-order volition to 'constitute his activity' is left unanswered.

Perhaps the thought is that second-order volitions are typically 'active' undertakings on our part, the result of mental effort. As Ekstrom (2005) points out, "a desire for *having another desire* (or for a certain desire to lead one to act, when or if one acts) is apparently not the sort of state that arrives unbidden" (p. 49). And, so the argument goes, whenever a desire is produced as a result of mental effort or deliberation in this way, the desire is necessarily 'his,' in a significant sense.

There are two problems with this suggestion though. First, although many of our second-order volitions are no doubt produced by mental effort, not all are. It is surely not a necessary truth that a person's second-order volitions are formed as a result of mental effort. Thus, simply because a person has a second-order volition, it does not follow that this constitutes his activity and so does not follow that the question of his identification with it cannot arise. Second, even if it were true that a person necessarily truly owns those parts of his mental life that come about as a result of effort, surely it is not a necessary truth that a person's first-order desires cannot be formed in this manner. Should a person form a first-order desire as a result of mental effort on his part, then this desire would necessarily be fully 'his,' a desire with which he necessarily identifies himself without it being the case that he desires to have it.

In his 1976 paper, 'Identification and externality,' Frankfurt broaches the new regress problem though he speaks of it in terms of 'attitudes' rather than desires and in

terms of 'internality' rather than identification and true ownership. He writes that according to his hierarchical account, an attitude's "internality will have to be accounted for by invoking a higher-order attitude – that is, an attitude toward an attitude" (p. 248). However:

> … the internality of this higher-order attitude will have to be accounted for in terms of an attitude of a still higher order," meaning that "an infinite regress will be generated by any attempt to account for internality or externality in terms of attitudes (p. 248).

Frankfurt suggests that the regress can be terminated by "making a particular kind of decision" (p. 250), though he says little about the content of such decision, calling the nature of decision "very obscure" (p. 251). He insists, though, that the making of a decision by the individual would terminate the regress because "decisions, unlike desires or attitudes, do not seem susceptible both to internality and to externality" (p. 251, footnote 3).

The suggestion from these admittedly sketchy remarks has the same problem as Frankfurt's earlier appeal to the way in which one's second-order volitions constitutes one's activity. In particular, we can ask *why* decisions are parts of a person's mental life that are necessarily fully his. And if the idea is that they are the products of mental effort, the result of some sort of 'work,' on our part, then, as with his remarks about second-order volitions, we avoid the regress only at the expense of giving up Frankfurt's central claim that identification must be understood in terms of a hierarchy of desires. This is because, as noted earlier, it seems conceptually possible for a person to form a first-order desire as a result of some effort on his part, without desiring to have that desire. I also

doubt that it is a necessary truth that decisions are made as a result of mental effort. If this claim is false, then making a decision will not necessarily terminate the regress.

In his 1992 paper, 'The Faintest passion,' Frankfurt suggests a different solution to this regress. He implies that a person identifies himself with his second-order volition if and only if he is 'satisfied' with that volition, where satisfaction "is a state of the entire psychic system – a state constituted just by the absence of any tendency or inclination to alter its condition" (1999: 104). This terminates the regress since satisfaction with one's second-order volition is a *negative* state, in the sense that it requires "no adoption of any cognitive, attitudinal, affective, or intentional stance" (p. 104) and so "does not entail an endless proliferation of higher orders and desires" (p. 105).

How can identification with a desire be explained in terms of an 'absence of any tendency or inclination to alter one's condition'? I think that what Frankfurt has in mind is that a person identifies himself with his second-order volition to the extent that he has no questions about whether or not he wants to have that volition. One problem with this idea is that the regress seems to be terminated arbitrarily, ceased simply because the person has no questions about his second-order volition. A second difficulty is that if what explains a person's identification with his second-order volition is the absence of any questions about it, then a person would seem to be able to identify himself with his motivating *first-order* desire if he has no questions about its pertinence from the second-order level. But this gives up on Frankfurt's idea of couching identification in terms of a hierarchy of desires.

Frankfurt is aware of this second objection to his solution. By way of reply, he writes:

> It is possible, of course, for someone to be satisfied with his first-order desire without in any way considering whether to endorse them. In that case, he is identified with those first-order desires. But insofar as his desires are utterly unreflective, he is to that extent not genuinely a person at all. He is merely a wanton (pp. 105-106).

Perhaps Frankfurt wants to claim that while we can say that a person who is 'satisfied' with a certain first-order desire may identify with it *in a sense*, he cannot do so in the sense required for the freedom for moral responsibility because, given his lack of second-order volitions, he is not a person but merely a wanton – in other words, he is not a being who can, in principle, exercise the freedom for moral responsibility. However, this puts pressure on being clear about Frankfurt's criterion for personhood. In his original 1971 paper, Frankfurt defines a person as a being who is "able to form … second-order desires" (p. 6), who has "the capacity for reflective self-evaluation that is manifested in the formation of second-order desires" (p. 7). The point implied here is that so long as an individual is *able* to form second-order volitions then he counts as a person, he counts as a being to whom 'freedom can be a problem,' to borrow Frankfurt's phrase. And it seems to me that there is no reason to think that a person might have no questions about whether or not he wants to have his motivating desire while being *able*, in the relevant sense, to form a second-order volition towards it, were he to *have* questions about it. The point, so far as 'satisfaction' is concerned, is that the individual is such that he has no *inclination* to form a second-order volition – not that, in principle, he *cannot* do so. Merely because a person's first-order desires are, on occasion,

'unreflective,' to use Frankfurt's phrase, does not entail that he is *not* a person, and so does *not* entail that he cannot, in principle, exercise the freedom for moral responsibility.

Alternatively, Frankfurt might mean to be claiming not that such an individual is a wanton *simpliciter*, but rather that he is a wanton *with respect to* those unreflective first-order desires. But I don't see how this really helps. Given Frankfurt's own account of what explains a person's identification in terms of satisfaction, it simply follows that if a person has a certain first-order desire and is such that he has no tendency or inclination within him to alter his condition, then he has identified himself with that desire and it is genuinely 'his' in the required sense. Nothing in Frankfurt's remarks above would seem to show this false.

Chapter Six: Undermining New Compatibilism I – Manipulation and Causal Responsibility

In the previous chapter, I evaluated Frankfurt's hierarchical account of the freedom required for moral responsibility. In this chapter, I want to turn my attention away from Frankfurt's particular account and look more generally at compatibilist views that try to capture moral responsibility's freedom without reference to alternative possibilities. These views can be thought of as an *indirect* attack on PAP. This is because such views, developed by so called *new compatibilists*, turn on developing accounts of the freedom needed for moral responsibility that do not rely on the freedom to do otherwise. If these conditions turn out to be sufficient to capture moral responsibility's freedom, then PAP would be false. With this challenge in mind, in this chapter I outline and evaluate two strategies in which PAP might be defended against these new compatibilist conceptions of freedom – the first involving manipulation cases and the second appealing to the difference between causal and moral responsibility. Despite my sympathies with the aims of these two strategies, I conclude that neither forms a decisive strike against the sufficiency of the new compatibilist conditions. A compelling argument undermining new compatibilism will have to be found elsewhere, a challenge I take up in chapter seven.

6.1 Self-expression and reasons-responsiveness.

Generally speaking, new compatibilist accounts of moral responsibility – those that do not make use of the freedom to do otherwise – can be divided into two kinds

depending on whether they emphasize self-expression or responsiveness to reasons as the key freedom-relevant feature. Frankfurt's (1971) is the first new compatibilist view of the self-expression type. On such a view, a person's being moved by a motivational element with which he identifies ensures that his behavior is a reflection of a fundamental part of his self; it ensures, we might say, that his action reflects 'who he really is,' or 'where he really stands' as a person.[25]

Other philosophers have followed Frankfurt in trying to capture moral responsibility's freedom in terms of actions that reflect certain privileged parts of the person's psyche. However, they have disagreed with him as to what part of a person's self his behavior must reflect. Gary Watson (1975), for instance, casts the freedom required for moral responsibility in terms of those action that are motivated by a person's value judgments as opposed to his 'mere' desires.[26] More recently, T. M. Scanlon (1998) ties free and responsible behavior to actions that reflect the individual's judgment-

---

25 Describing his view about the relationship between freedom and moral responsibility, Frankfurt (1971) writes, "It is a mistake, however, to believe that someone acts freely only when he is free to do whatever he wants or that he acts of his own free will only if his will is free. Suppose that a person has done what he wanted to do, that he did it because he wanted to do it, and that the will by which he was moved when he did it was his will because it was the will he wanted. Then he did it freely and of his own free will … Under these conditions, it is quite irrelevant to the evaluation of his moral responsibility to inquire whether the alternatives that he opted against were actually available to him" (p. 19).

26 Watson (1975) describes a person's free actions – and, by implication, his morally responsible actions – as those that "flow from his evaluational system" (p. 216). This system comprises "that set of considerations which, when combined with his factual beliefs (and probability estimates) yield judgments of the form: the thing for me to do in these circumstances, all things considered, is *a*" (p. 215). The set of considerations to which Watson is referring include his values, defined as "those principles and ends which he – in a cool and non-self-deceptive moment – articulates as definitive of the good, fulfilling and defensible life" (p. 215).

sensitive attitudes, while Angela Smith (2005, 2008) speaks of such actions as being those that reflect the person's rational judgments.[27]

Despite the differences in the details of their views, in my opinion Watson, Scanlon, and Smith share with Frankfurt the basic idea of understanding the freedom required for moral responsibility in terms of, what I call, the *freedom of self-expression* rather than the freedom to do otherwise. On the self-expression model, a person acts with the freedom required for moral responsibility to the extent that his action reflects the part of his self that is most fundamental to who he really is, as a person. While Frankfurt, Watson, Scanlon, and Smith disagree as to what constitutes the part of a person's self that is really 'him' – whether it is his identified desires, his considered values, his judgment-sensitive attitudes, or his rational judgments – the important point is that they agree that self-expression is all the freedom required for moral responsibility. And if this were true, then PAP would be undermined. This is because in order for a person's action to reflect his 'true' self, it is not the case that he must have been free to do otherwise.

Other compatibilists who share Frankfurt's project of capturing the freedom required for moral responsibility without reference to alternative possibilities have offered a different conception of freedom from the freedom of self expression, one I shall call the *freedom to respond to reasons*. According to the reasons-responsiveness picture, a person acts with the freedom required for moral responsibility to the extent that he

---

[27] Scanlon (1998) claims that knowing whether a person is morally responsible for his behavior consists in "determining whether a given action did or did not reflect that agent's judgment-sensitive attitudes" (p. 290). Smith (2008) suggests that "to say that an agent is morally responsible for something, on this view, is to say that that thing reflects her rational judgment in a way that makes it appropriate, in principle, to ask her to defend or justify it" (p. 369).

regulates his behavior by moral reasons. John Martin Fischer and Mark Ravizza (1998) and R. Jay Wallace (1994), among others, defend views of this type. According to Fischer and Ravizza, whose view has been most influential, a person acts with the freedom required for moral responsibility if and only if the process that led to his action was responsive to moral reasons in the sense that, holding that process fixed, in some possible circumstance in which there were sufficient moral reasons to do otherwise, the individual would have recognized these reasons in an understandable way and would have acted differently for at least one of them.[28] These reasons-responsiveness conditions threaten to undermine PAP in just the same way as self-disclosure views, since regulating one's behavior by reasons does not require the freedom to do otherwise.[29] So, if reasons-responsiveness is all the freedom needed for moral responsibility, then PAP would be false.

How might the sufficiency of these new compatibilist conditions be challenged? How, in other words, can PAP be defended in the face of them? In a context in which what is at issue is whether or not moral responsibility requires the freedom to do

---

28 Describing this condition, Fischer and Ravizza (1998) write, "a mechanism of kind *K* is moderately responsive to reasons to the extent that, holding fixed the operation of a *K*-type mechanism, the agent would *recognize* reasons (some of which are moral) in such a way as to give rise to an understandable pattern (from a viewpoint of a third party who understands the agent's values and beliefs), and would *react* to at least one sufficient reason to do otherwise (in some possible scenario)" (pp. 243-244). (By 'sufficient' reason, Fischer and Ravizza mean a reason that is all-things-considered his best or strongest reason for action.) Fischer and Ravizza have a *further* freedom–relevant requirement for moral responsibility, namely that a person must 'take responsibility' for the process that led to his action. This additional condition makes no difference to the force of my criticism of their view.

29 One might argue here that regulating one's behavior by moral reasons *does* require some sort of freedom to do otherwise. However, I set this concern aside in what follows.

otherwise, it would be straightforwardly question-begging against those who reject PAP to argue that these compatibilist views are not sufficient to capture the freedom required for moral responsibility precisely because they make no use of alternative possibilities. With this in mind, I want to outline and evaluate two prominent lines of argument in the literature that are designed to challenge the sufficiency of the new compatibilist conditions without falling into this question-begging charge. One has to do with manipulation, while the other concerns the difference between causal and moral responsibility.

6.2 Manipulation arguments.

Manipulation arguments, which have recently been defended by Robert Kane (1996) and Derk Pereboom (2001) among others, turn on the claim that people can be covertly manipulated by neuroscientists into satisfying the relevant compatibilist conditions. However, being subject to this sort of manipulation is intuitively at odds with free and responsible action, and so these compatibilist conditions cannot be sufficient to capture the freedom needed for moral responsibility.

To illustrate, consider an example derived from Pereboom's (2001) well-known case of Professor Plum. Plum murders Ms. White for the sake of some personal advantage. Causal determinism is true and Plum satisfies the new compatibilist conditions for the freedom required for responsibility that we have discussed. His desire to kill White is his will because it was the will he wanted (Frankfurt's condition). His action reflects attitudes that are judgment-sensitive (Scanlon), judgments that are both rational (Smith) and represent Plum's values (Watson). Finally, his behavior is

responsive to reasons in the sense that holding the process that led to his action fixed, in some possible circumstance in which there were sufficient moral reasons to do otherwise, Plum would have recognized these reasons in an understandable way and would have acted differently for at least one of them (Fischer and Ravizza).

However, unbeknownst to Plum, his satisfaction of these conditions was entirely due to the covert behavior of a team of neuroscientists who wanted to be sure that he would kill White. They covertly manipulated him by interfering with the neural connections in his brain ensuring that he met these conditions. This by itself implies that these compatibilist conditions are not sufficient to capture moral responsibility's freedom. Yet, in the face of the possible compatibilist reply that it is one thing for neuroscientists to manipulate a person but another thing for him to act in the 'normal,' unimpaired way, Kane, Pereboom, and others, take the argument further. They claim that there are no morally relevant differences between a person who is determined by neuroscientists into meeting these conditions and one who is determined by the laws of nature and facts about the past into satisfying them. If this is true, then people's non-responsibility in the manipulation case would transfer to any instance in which they satisfy the compatibilist conditions as a result of 'mere' causal determinism.

The soundness of this kind of argument has recently been challenged by compatibilists. In the next section of the chapter, I will assess this debate. Before that, however, I want to consider a dialectical issue about using manipulation cases to argue that it is the absence of the freedom to do otherwise in the manipulation example that explains why Plum does not act freely and responsibly.

Manipulation arguments are standardly used by incompatibilists to argue that the new compatibilist conditions are not sufficient to capture the freedom required for moral responsibility. However, what is interesting about these arguments is that the kinds of incompatibilists that develop them are not usually those who believe that PAP is true. In fact, although incompatibilists use manipulation arguments as a way to undermine the sufficiency of the compatibilist conditions, they do not usually use them to support the claim that what is missing in these compatibilist accounts is the (incompatibilist) freedom to do otherwise. Instead, Kane, Pereboom, and other so called *source incompatibilists*, use them to support their claim that "an action is free in the sense required for moral responsibility only if it is not produced by a deterministic process that traces back to causal factors beyond the agent's control" (Pereboom, 2001: 3). (Kane endorses a similar principle.) On this line of thought, what explains our intuition that the manipulated individual does not act freely and responsibly is not that he lacks alternative possibilities *per se* but rather that his actions trace back to deterministic factors beyond his control, namely, the intentions and actions of the neuroscientists. In other words, according to source incompatibilists, it is the fact that Plum is not the true source of his behavior rather than his lacking the freedom to do otherwise that explains why we have the intuition that he does not act freely and responsibly.

This use of the manipulation argument potentially undermines my defense of PAP. After all, if sound, the argument would apparently undermine the sufficiency of the new compatibilist conditions at the expense of conceding that the freedom to do otherwise is not what is missing from these accounts and so not something that is necessary for moral responsibility. However, I do not think that this is a serious

challenge to my defense of PAP. Even if manipulation arguments were sound, I doubt whether this would be a strike against PAP because it is far from clear that manipulation cases *only* support a source-based version of incompatibilism. After all, our intuition that the manipulated person does not act freely and responsibly might be best explained by the fact that the manipulation impedes his ability to do otherwise. According to this thought, it is because manipulated Plum is unable to do otherwise, rather than the fact that his behavior traces back to the intentions of the neuroscientists, that explains why we regard him as lacking freedom and responsibility.

However, I do not think that this dialectical issue needs to be settled because – as I now want to explain – compatibilists have developed a response to manipulation arguments that, at least in my view, challenges the decisiveness of these arguments. If manipulation arguments are not decisive, then it would be a moot point whether or not they would succeed in undermining the new compatibilist conditions at the expense of conceding that alternative possibilities are not required for moral responsibility.

6.3 The Dilemma response to manipulation arguments.

The soundness of manipulation arguments has recently been challenged by compatibilists. Some reject the claim that there are no morally relevant differences between a manipulated individual and an unimpaired person (what Michael McKenna (2008b), following Kane (1996), calls a 'soft-line' reply). Others argue that when manipulated in a particular kind of way it is plausible to think that a manipulated individual *does* act freely and responsibly after all (what McKenna (2008b) terms the 'hard-line' reply). One way to illustrate these two lines of reply is in terms of a dilemma.

I call this the *dilemma response* to the manipulation arguments. Frankfurt (1975) originally developed the dilemma but it has been subsequently taken up, in one form or another, by Lynne Rudder Baker (2006), Fischer (2006), McKenna (2008b), and Watson (1999). The dilemma turns on whether or not the manipulation is moment-to-moment or a one-time occurrence early in the individual's life.

Thinking about the Plum example, one question we can ask is how the neuroscientists go about manipulating Plum into meeting the compatibilist conditions in the first place. On the one hand – the *first* horn of the dilemma compatibilists pose in the face of these examples – we can imagine that neuroscientists manipulate Plum in a *moment-to-moment* or on-going basis. In this case, they program "him to undertake the process of reasoning by which his desires are brought about and modified – directly producing his every state from moment to moment" (Pereboom, 2001: 112-113). They do this by "pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning to be rationally egoistic" (p. 113) which, in the situation, leads him to kill White.

However, these compatibilists claim that this kind of programming does not undermine the sufficiency of the compatibilist conditions. They argue that what explains why Plum does not act freely and responsibly is not that the compatibilist conditions are insufficient but that the moment-to-moment manipulation of the neuroscientists undoes the presence of background conditions of agency that all sides agree are needed for free and responsible action. For instance, compatibilists and incompatibilists generally agree that in order for a person to act freely and responsibly it must be the case that he has the capacity to step back and critically evaluate his mental states. However, the on-going

manipulation to which Plum is subject would prevent him from being able to assess his desires, motives, and values in this way since the neuroscientists would be directly inducing new states into him in a moment-to-moment basis. (Frankfurt [1975] signals this by speaking of the neuroscientists' manipulation as rendering a person little more than a "marionette" [p. 120].) So, by this line of reasoning, Plum's non-responsibility in the local manipulation scenario is attributable to his lack of basic agential capacities, capacities that are needed in order for him to act freely and responsibly. Importantly, these capacities *would* be present were the action deterministically caused in the normal, unimpaired way, so we can mark a morally relevant difference between determination by neuroscientists and determination by natural causes.

On the other hand – the *second* horn of the compatibilist defense – we can conceive of Plum being manipulated not in a 'local,' on-going way but, instead, as a one-time occurrence earlier in his life. At an earlier point in his life, the neuroscientists' "programmed him to weigh reasons for action so that he is often but not exclusively rationally egoistic, with the result that in the circumstances in which he now finds himself," he kills White while meeting the compatibilist conditions (Pereboom, 2001: 114-115). The fact that the neuroscientists' programming is temporally remote rather than moment-to-moment removes the worries from the previous case about whether or not Plum possesses the right background conditions of agency that are necessary for free and responsible action. In this kind of situation, Plum presumably retains the capacity to reflect on his desires and values since the manipulators would not be constantly inducing new states into him.

However, in these circumstances, compatibilists insist that it is no longer as clear that Plum does not act freely and responsibly. As McKenna (2005b) puts it:

> … once the manipulation is so qualified that all an agent's current time-slice compatibilist-friendly structures are properly installed through a process of manipulation, then the role of the manipulator begins to shrink into the background; we are simply left with a normal person who happened to be brought into existence in a very peculiar manner (p. 217).

So long as the manipulation is remote enough not to impair the individual's powers of critical assessment then, as Frankfurt (1975) urges, he "may become morally responsible, assuming that he is suitably programmed" (p. 120). (Baker [2006], McKenna [2008b], and Watson [1999] have expressed agreement with Frankfurt on this point.) What are we to make of this?

With respect to the remote Plum case, my intuitions clash with Frankfurt and the other compatibilists. Intuitively, it seems to me, we would *not* be inclined to think that Plum acted freely and responsibly in killing White were we to discover that he was the victim of such manipulation, even if the manipulation was temporally remote. In fact, I would argue that so long as the manipulation impaired Plum's ability to do otherwise, then whether or not it was local or remote is irrelevant to the issue of his responsibility.

However, I realize that not everyone will share my intuition. For those whose intuitions are not settled, compatibilists taking this apparently 'hard line' could make the claim that Plum acts freely and responsibly more plausible by drawing attention to the similarities between the remote manipulation case and those of a normal upbringing. After all, we are all to some degree products of our circumstances, our family

environment, our genetic inheritance, and so on. Yet, we do not tend to think that these facts undermine our freedom and responsibility, so why should things be different in the case of someone whose conduct is fashioned by these neuroscientists rather than by the more usual formative circumstances?

The growing tendency among compatibilists to question the intuition that Plum does not act freely and responsibly so long as the manipulation does not undermine his basic agency has led some to suggest that manipulation arguments fall into a "dialectical stalemate" (McKenna, 2008b: 145) implying that they are "unlikely to be a profitable interaction between compatibilists and incompatibilists" (King, 2009: 5). While talk of a stalemate is perhaps too strong – as Pereboom (2008) points out, "that term connotes the idea that the game is over and there are no further moves to be made" (p. 169) – and while I myself do not share the compatibilist intuition that Plum acts freely and responsibly, I do think that the compatibilist dilemma response constitutes an effective challenge to the *decisiveness* of manipulation arguments. In response to this, I want to look at other ways in which the new compatibilist conditions might be undermined in a way that does not fall into a clash of intuitions about particular cases. In other words, I want to ask whether there is a way to move the manipulation argument forward in the face of this apparent intuitive clash. To do this, I outline and evaluate a different argument against the new compatibilist conditions, one that has its roots in some work by Susan Wolf (1990). I call it the causal responsibility argument.

6.4 Causal and moral responsibility.

Should a compatibilist question the intuition that the remotely manipulated individual does not act freely and responsibly, I think that there is a possible line of response to be found in some remarks by Wolf (1990). We can argue that while the fact that a person's action reflects his 'real' self or is responsive to reasons does seem to capture the idea that he is a genuine *cause* or *author* of what he does, this only gives rise to a kind of responsibility that is causal rather than moral in character, and the question of a person's moral responsibility is different in kind from the casual question of whether or not the person brought about or 'authored' his behavior. Wolf describes the responsibility here as being 'superficial' rather than 'deep' in kind, moral responsibility having a quality of depth not captured by these compatibilist views.

Describing 'real' self views, Wolf (1990) argues that acknowledging that someone's action reflects his 'real' self helps "identify the individual as playing a causal role that, relative to the interests and expectations provided by the context, is of special importance to the explanation of that event" (p. 40). However, she claims that "when we hold an individual morally responsible for some event, we are doing more than identifying her particularly crucial role in the causal series that brings about the event in question. We are regarding her as a fit subject for credit or discredit on the basis of the role she plays" (pp. 40-41).

Wolf tries to motivate the difference between causal and moral responsibility as follows:

> … it is intelligible to wonder whether a person is deeply responsible for an action even after we have removed all doubt that she really did perform that action. We

> can coherently acknowledge that a person really did play a relevantly crucial role in bringing a very good or very bad event about, and yet be uncertain about whether the person deserves to be praised or blamed for it (p. 41).

As I would put it, and broadening the criticism to apply to reasons-responsiveness views as well as to 'real' self theories, the causal question of being sure *that* someone did something is different in character from the moral question of whether the person deserves praise or blame *for* doing that thing. So, even if we were sure *that* the person really performed the action – because, for instance, it reflects who he really is or is responsive to reasons – it would still be an open question whether he is blameworthy or praiseworthy *for* behaving as he did. Knowing that people genuinely author their actions does not settle the matter as to whether they are morally responsible for them.

I think that this argument – what I call the *causal responsibility argument* – is very suggestive. However, as I will explain shortly, it too falls short of being decisive because compatibilists have a response open to them that blunts much of its force. Before I outline this response, though, I want to first consider a different line of defense new compatibilists might make, one due to Watson (1996). In my view, Watson's defense is not compelling and new compatibilists would do better to adopt the line of reply I go on to suggest.

6.5 Watson's two faces of responsibility.

Against Wolf, Watson (1996) develops two main points. First, he argues that when we properly understand 'real' self views – at least the type he favors – we see that they *do* give rise to a kind of responsibility that is moral and not simply causal in

character. However, in his second point, he concedes to Wolf that the relationship between self-expression and moral responsibility is not quite so simple. He argues that there is more than one kind – or 'face,' as he puts it – of moral responsibility. While self-expression is sufficient to capture the freedom required for one kind of moral responsibility, what he calls 'attributability,' there is a *further* kind of moral responsibility that he calls 'accountability' whose conditions of freedom outstrip self-expression.

With respect to this first point, Watson argues that 'real' self views *do* capture a genuinely moral kind of responsibility. In his view, it is a person's considered values or "fundamental evaluative orientation" (p. 234) that constitutes his 'real' self or who he really is as a person. Because of this, when people's actions reflects their values, they can be morally evaluated – and hence morally responsible – in ways that bear upon their evaluative orientation. In particular, they can be morally responsible in terms of what their actions reveal about the kind of moral agents they are.[30] Watson gives an example of a thief stealing some books (pp. 230-231). If the thief's behavior is reflective of his values, then he can be evaluated in ways that bear upon his value judgments. In this instance, by deciding to steal, the thief shows himself to be a bad, callous, or perhaps evil person. "These appraisals," Watson suggests, "concern the agent's excellences and faults – or virtues and vices – as manifested in thought and action … [and so we can say that] such judgments are made from the *aretaic perspective*" (p. 231).

---

30 It is unclear whether Frankfurt could make use of Watson's defense here because, on Frankfurt's view, a person's identified desires – which comprise his 'real' self – have no necessary connection to his value judgments or conception of the good (see Frankfurt 1971: 13, footnote 6).

However, in his second main point, Watson argues that not all moral responsibility corresponds to this kind of aretaic or attributability responsibility. There is a further kind – or 'face' – of moral responsibility called 'accountability' which has to do with the practice of "holding people morally accountable" (p. 230) for their behavior. To hold someone morally accountable for his action is not merely to evaluate what his actions reveal about the kind of person he is; rather, it is to respond to him in *further* ways that are characteristic of holding people responsible – subjecting them to sanctions in the negative case (like censure or punishment) or provided them with benefits in the positive case (like public admiration or reward). Furthermore, the kind of freedom needed for people to be accountable for what they do is not met when their actions reflect their 'real' selves. In addition, Watson implies that a PAP-type condition applies to accountability in the sense that, at least in the negative case, people must have had a reasonable opportunity to avoid incurring sanctions in order for it to be fair for them to receive such treatment.[31]

Is Watson's appeal to two faces of responsibility a good defense against the causal responsibility argument? I do not think so. The first question I would ask is whether attributability should be thought of as a kind of moral responsibility at all. On the one hand, Watson is surely right that a when a person's action reflects his values, he shows himself to be a particular kind of moral agent, and so evaluations of him will have a moral quality beyond mere causal assessment. As Wallace (1994) puts it, "assessments

---

31 In more recent work, Watson (2004) acknowledges that the PAP-type condition required for accountability "creates a foothold for incompatibilist doubts" (p. 10). It is unclear from his remarks whether Watson sees self-expression as *necessary* for accountability. I follow Smith (2008) in suggesting that he does.

of people as responsible, in the sense of being autonomously self-revealing in their actions, clearly go beyond mere descriptions of causal responsibility, and so may be said to have a quality of depth" (p. 53).

In fact, I concede to Watson that some philosophers have not recognized that this point should be common ground. Wolf (1990) herself argues that evaluating people in this way is no different in kind from the assessments we make of "earthquakes, defective tires, and broken machines" (p. 42). She makes this claim on the basis that we can attribute the city's destruction to the earthquake or the car's failure to the damaged tire in just the same way that the actions of people can be genuinely attributed to them. However, her argument is surely wrong. Full-fledged moral agents open themselves up to certain forms of moral criticism and evaluation when their actions reveal the kind of moral agents they are that are simply not applicable to earthquakes, tires, or even nonhuman animals and young children. However, what *is* open to question and should *not* be common ground is whether these evaluations and this responsibility should be classed as a kind of moral responsibility. Let me motivate this doubt.

All agree that there is a deep conceptual tie between moral responsibility on the one hand and moral praise and blame on the other. Many claim, for instance, that to say a person is morally responsible for what he does is to say that he deserves moral praise or blame, in one form or another, for what he does.[32] What is striking, though, is that judgments about the kind of moral agents people are – those judgments licensed by Watson's condition of attributability – are different from, and do not seem describable in terms of, moral praise or blame. After all, it is one thing to judge a thief to be a callous,

selfish, or evil person in virtue of *what* he did; it is another thing, however, to judge him blameworthy *for* acting as he did. This difference reflects a natural picture according to which not all the moral assessments we can make of people bear on the question of their moral responsibility. On this picture, the assessments that *do* bear on this question are those pertaining to a person's blameworthiness or praiseworthiness. And since the assessments licensed by 'real' self views are not reflective of moral praise or blame, they should not be thought of as judgments of moral responsibility.

Watson seems to be aware of this objection since he argues that 'real' self views *do* give rise to assessments that are a kind of moral praise and blame. "In one way," he writes, "to blame (morally) is to attribute something to a (moral) fault in the agent; therefore, to call conduct shoddy *is* to blame the agent" (pp. 230-231). As he later puts it, "the aretaic perspective is a source of blaming judgments in one plain sense: judgments that the agent's conduct was faulty in some way. If the fault is moral, so is the blame" (p. 238). However, I do not share Watson's intuition that to call a person's conduct shoddy is to blame him in any moral sense. At most, I see how one might argue that in calling someone's behavior shoddy we might, in part, be judging that the action is a reflection of a moral failing on the individual's part and he is, in this sense, 'to blame' for it. However, this is simply a *causal* use of the term blame. To say that the person is 'to blame' for the action in this context is simply to say that *he* is the one to whom the failing can be attributed, as an exercise of *his* moral agency. As such, any sense in which we might be tempted to speak of the evaluation of a person's moral character as being a kind

---

[32] See, for example, Fischer and Ravizza (1998), Pereboom (2001), and Wolf (1990).

of blame or praise is best explained as being a *causal* sense of blame or praise rather than the moral kind or praise and blame that is at issue in contexts of moral responsibility.

Watson could, of course, just *stipulate* that judgments of attributability are describable in terms of moral praise or blame. However, it is unclear how far this can take him. Stipulating that assessments of a person's moral character are kinds of moral praise or blame leaves little room for the possibility of moral evaluation of a person that is *not* a type of moral praise or blame. This is surely not right. We want to know how to divide up the moral assessments we make of people into those that pertain to moral praise and blame and those that do not. We do not want to collapse this distinction entirely.[33]

So far, I have cast doubt on Watson's claim that attributability, the sort of responsibility that 'real' self views apparently undergird, should be thought of as a kind of moral responsibility. However, I now want to turn to a different criticism of Watson's defense. In particular, I would like to argue that even if we were to grant for the sake of argument that attributability does capture a genuinely moral kind of responsibility, there

---

33 An alternative response Watson and his defenders could make in the face of my criticism is to reject the conceptual link that philosophers have generally drawn between moral responsibility on the one hand and moral praise and blame on the other. If we relaxed this link, and broadened the scope of evaluations a person could be open to that bear on the question of his moral responsibility, then assessments of his character could be classed as judgments of moral responsibility *whether or not* they could be interpreted as forms of moral praise or blame. Pereboom (Fischer, Kane, Pereboom & Vargas, 2007; Pereboom, 2008), for instance, follows Watson in speaking of there being different 'senses' or 'notions' of moral responsibility, but unlike Watson he implies that some of these senses of responsibility do *not* give rise to moral praise or blame. One of these is, what he calls, the 'legitimately called to improvement' sense, a kind of responsibility that closely matches what Watson has in mind with his idea of attributability. A person is morally responsible in this sense if, among other things, she can be legitimately open to critical evaluation of "what her decisions and actions indicate about her moral character" (Fischer, Kane, Pereboom & Vargas, 2007: 86). In response, I find it deeply unattractive,

are *still* reasons to think that his defense of 'real' self views is unsatisfactory. This is because Watson admits that it is the *other* face of responsibility, *accountability*, rather than attributability that captures the "ordinary, full-fledged concept of moral responsibility" (p. 243).[34] This is significant because new compatibilists have generally taken themselves to be after the same thing as PAP-defenders. That is, both parties have been trying to articulate the kind of freedom with which people must act if they are to be morally responsible for their behavior in the ordinary, traditional sense of that term. However, by conceding that 'real' self views do not capture the freedom needed for this ordinary kind of moral responsibility, even those sympathetic to such views should find Watson's defense less than satisfactory.

Finally, Watson tries to motivate his argument that attributability is a genuine kind of moral responsibility by arguing that it is "central to ethical life" (p. 243). This is because making assessments about the kind of moral agents people are helps us settle our

---

and a high price to pay to defend a theory, to give up the tight link between moral responsibility and praiseworthiness and blameworthiness.

34 Pereboom (Fischer, Kane, Pereboom & Vargas, 2007) expresses similar sentiments. He claims that while the idea of moral responsibility as centered upon evaluations of a person's moral character, the 'legitimately called to improvement' sense, for instance, "may be a *bona fide* sense of moral responsibility, it is not the one at issue in the free will debate." This is because "incompatibilists would not find our being morally responsible in this sense to be even *prima facie* incompatible with determinism. The notion that incompatibilists do claim to be at odds with determinism is the one defined in terms of basic desert" (p. 86). While I am sympathetic to Pereboom's conclusion here, I do not think that he has quite captured the difficulty. The point is not, as he suggests, that responsibility-as-character-assessment is not the sense of moral responsibility at issue because it is not defined in terms of basic desert *simpliciter*. That notion may *well* be defined in terms of basic desert, in terms of whether or not a person deserves to be judged as being a particular kind of moral agent in virtue of his action. Rather, the reason this notion is not the central one at issue is because it is not defined in terms of whether or not a person deserves *to be judged morally praiseworthy or blameworthy* for what he has done.

own convictions about "living a good human life" (p. 243) and, as he puts it in more recent work, they help us better form ideas about "the possible models of human achievement and failure" (2004: 10). I am inclined to agree with Watson that our capacity to evaluate the moral character of ourselves and our fellow human beings is deeply significant. However, I fail to see how this fact makes it any more plausible that these kinds of assessments should be thought of as bearing on the question of a person's moral responsibility. Why not simply insist that they are significant to our moral lives and be done with it?

6.6 The Epistemic condition of moral responsibility.

In the previous section, I argued that Wolf's argument is not undermined by Watson's appeal to two faces of responsibility. However, despite my sympathies with its aims, I do not think that the causal responsibility argument is ultimately decisive. Just as compatibilists had a compelling response to the manipulation argument available to them, I also believe that they have a response to the causal responsibility argument that blunts its apparent force. Let me explain.

Compatibilists and incompatibilists generally agree that acting with a certain kind of freedom is not the sole consideration in determining a person's moral responsibility. In order to assess whether an individual is morally responsible for what he does, all sides agree that we should not only enquire about the freedom with which he acts. We also need to ask whether he satisfies the relevant *epistemic* conditions for moral responsibility. Did he know what he was doing and what effects his actions would likely have? If not, *should* he have known this? Was his ignorance culpable? It is common ground that as

well as knowing the freedom with which someone acts we also need to know details about the knowledge and beliefs he has when acting in order to assess whether or not he deserves praise or blame for what he does.

However, bringing to light the epistemic component of moral responsibility reveals a way for compatibilists to challenge the decisiveness of the causal responsibility argument. As Wolf presented it, the argument is premised on the claim that the new compatibilist conditions are not sufficient to capture moral responsibility's freedom because even if we are sure that the individual has met these conditions and so is causally responsible, we can still raise the further question of whether he is *morally* responsible for what he does. Yet compatibilists should insist that the fact that the moral question remains open under these circumstances is exactly what we should expect. So long as we do not yet have details about whether the person satisfies the relevant epistemic condition for moral responsibility (whatever this is), simply knowing the freedom with which he acts would not be enough – as both compatibilists and incompatibilists agree – to settle the question of whether he deserves praise or blame for his behavior.

The mere fact that the moral question is open under these circumstances does not, therefore, show that the acting so as to reflect one's 'real' self or in a way that is responsive to reasons is only sufficient to give rise to a kind of responsibility that is causal rather than moral in character. The error in Wolf's argument is to fail to appreciate that moral responsibility has *two* pertinent conditions – one having to do with freedom, the other concerning the individual's epistemic state – and it is only when we have details about *both* of these conditions that the question of a person's moral responsibility should be closed.

Chapter Seven: Undermining New Compatibilism II – Blame, Demands, and Authorship

In the previous chapter, I argued that two prominent arguments in the literature – the manipulation argument and the causal responsibility argument – do not decisively undermine the sufficiency of the new compatibilist conditions. If my claims in the previous chapter are correct, then those who are sympathetic to PAP are faced with the question of how *else* the principle might be defended in the face of the new compatibilist challenge. In this chapter, I develop a different way of defending PAP that brings to light and questions a traditional assumption about the relationship between freedom and moral responsibility.

7.1 Blame, expectations, and demands.

I want to begin by outlining and defending an argument for PAP's truth that draws on some recent remarks by David Widerker (2005). I then go on in later sections of the chapter to show that, if sound, this Widerker-style argument has profound dialectical implications. In particular, if sound, it represents a significant shift in our understanding of the relationship between freedom and moral responsibility, something that has so far gone unnoticed.

Several philosophers have recently drawn attention to the apparent link between moral blame and moral expectations or demands. They have pointed out that there is an intuitive connection between blaming people on the one hand and expecting or, perhaps,

demanding things of them and their behavior on the other. To cite one instance of this, consider Michael McKenna's (2008a) claim that:

> When we blame a person for a moral wrong, a clear implication is that our moral charge includes the demand that the person not have done that, that the person have acted as morality requires (p. 783).

The particular question I want to take up is how we can best capture the apparent relationship between these concepts. Widerker (2005) suggests that the key idea here is that a person's blameworthiness rests, in part, on it being reasonable for those in a position to do so to expect of him that he not have acted as he did. He captures this insight with the following principle, which he calls the *principle of alternative expectations (PAE):*

> An agent *S* is morally blameworthy for doing *A* only if in the circumstances it would be morally reasonable to expect of *S* not to have done *A*.[35]

However, the trouble with this principle is that the idea of expecting people to act in certain ways but not in others, in the sense Widerker intends, is not obvious and needs clarification. Clearly, the sense of expectation he has in mind is not that of prediction. That is, to have expected a person not to have behaved as he did, in the peculiar sense Widerker is after, is not for us to have made a *prediction* about how we believe he would likely have acted. Rather, for us to have expected a person not to have acted as he did is for us to have *demanded* or, perhaps, *insisted* that he not behave that way. To use some

---

35 By it being 'morally reasonable' to expect the person not to have acted as he did, Widerker points out that he means morally reasonable "for someone who is morally competent and knows all the relevant non-moral facts pertaining to the situation the agent is in" (p. 297, footnote 20).

examples, the sense of expectation at issue here is that we use when, in situations in which people are being dishonest, we expect or demand that they tell the truth, or in circumstances where people are keeping things that do not belong to them, we demand that they return what they owe.

To avoid confusion about the notion of expectation that is relevant when it comes to a person's blameworthiness, I want to rephrase Widerker's principle in terms of demands rather than expectations. For, as I have suggested, it seems to me that by speaking of expecting people to act in certain ways but not in others, Widerker wants to convey the thought that blameworthiness is linked to the reasonableness of *demanding* that people behave in these ways. Let us call the revised principle the *principle of alternative demands (PAD):*

> An agent *S* is morally blameworthy for doing *A* only if in the circumstances it would be reasonable for those in a position to do so to demand that *S* not have done *A*.

Having clarified the relevant notion of expectation by offering a revised version of Widerker's original principle, I will now explain this principle's importance. As Widerker points out, this kind of principle is significant because, if true, it can be used to support the claim that moral responsibility, at least as it applies to blameworthiness, requires alternative possibilities. It would therefore demonstrate that the new compatibilist conditions which do not make use of the freedom to do otherwise are not sufficient to capture the freedom required for moral blame. To see this, consider the following claim. If a person was *not* free to do otherwise when performing a wrong action, then it would *not* be reasonable to expect or demand that he not have acted as he

did. For if the person lacked alternative possibilities, then to demand this of him would be to demand that he have done something "impossible," as Widerker (p. 297) puts it, and have done something that it was not within his power to do. Yet, it is surely unreasonable or unfair to expect or demand that an individual have done something if this thing – that is, not acting as he did – was not something that was within his power to do.

According to this line of argument, in order for us to reasonably demand that a person not have behaved as he did, it must be the case that he had alternative possibilities. Furthermore, if the reasonableness, or fairness, of such a demand is a necessary condition for justified blame as PAD states, then we have an argument, against Frankfurt, Fischer, and the other new compatibilists, that blameworthiness requires the freedom to do otherwise after all.

This argument – which I call the *reasonable demand argument* – is provocative. In the next section, I want to defend it against some criticisms as a way of clarifying the theoretical grounding on which it rests.

7.2 Defending and motivating the argument.

I see two main ways in which the reasonable demand argument might be challenged. First, critics might argue that, contrary to one of the argument's premises, it is *not* unreasonable or unfair to demand that people who lacked alternative possibilities not have acted as they did. Perhaps, so the thought goes, it *can* be reasonable to make demands of how people behave even if it is not within their power to meet these demands. In other words, to retool a more familiar slogan about 'ought,' the critic is challenging the claim that 'demand-implies-can'.

But I doubt that this response has much force. In fact, I am inclined to think that the only reason that critics would suggest it or claim to find it plausible is if they already had a prior commitment to PAP's falsity. After all, it seems deeply plausible to think that such demands would *not* be reasonable under these circumstances. How can it be reasonable or fair to demand that someone not have acted as he did if this was not something that it was within his power to do? At any rate, I want to set this criticism aside and assess a second, more promising line of attack.

A second line that critics could take to undermine the argument is to challenge the truth of the principle on which it rests – that is, to reject PAD itself. In a recent paper, Justin Capes (2009) offers what he takes to be a criticism of Widerker's original principle, PAE, which, by extension, could also be used to challenge my revised version of it, PAD. He writes:

> The Frankfurt-defender, I am suggesting, has been given no reason to accept this claim [i.e., Widerker's claim that a person is blameworthy only if it would be reasonable to expect him not to have acted as he did]. Why must we expect [or demand] a person to do otherwise in order to disapprove morally of his behavior? It seems entirely possible to disapprove of a person's behavior and indeed to hold that person blameworthy for what she has done without there being an expectation [or demand] on our part that she not have behaved in that way. To be sure, we will no doubt wish that the person not have behaved as she did, and we may judge that the person ought not to have behaved that way, but … this needn't involve an expectation or demand on our part that the person not behave as she did (p. 15).

Capes is attempting to question the truth of Widerker's original principle – and, by extension, my revised version of his principle – on the grounds that we can blame people for what they do without apparently making any expectations or demands of them at all. But once we bring to bear a simple distinction we see that Capes' criticism is off-target and is no threat to the claim that blameworthiness requires the reasonableness of an expectation or demand. He mistakenly equates Widerker's principle with the stronger claim that blame *involves* an expectation or demand that the person not have acted as he did.[36] Yet these are distinct claims. In fact, the claim that blameworthiness requires that it must be reasonable to expect or demand that someone not behave as he did does not entail that his blameworthiness *involves* such an expectation or demand.

Pointing out that a person can appear blameworthy without an expectation or demand being made of him is *not*, therefore, a good criticism of either Widerker's original principle or my revised version. In fact, Widerker can admit that a reason for preferring his principle over the stronger claim that Capes mistakenly attributes to him – the claim that blame *involves* an expectation or demand – is precisely *because* it seems far more conceivable to blame an individual without demanding or expecting anything of him than it does to conceive of blaming him without it being reasonable *to* form a demand or expectation of him should we wish to.

While Capes' criticism misses the mark, I do think that there is room to question Widerker's own account of what explains the *attractiveness* of his principle (and, by extension, what would explain the appeal of my revised version of it). At one point in his

---

36 The stronger claim suggested in Capes' remarks is this: The blameworthiness of an agent *S* for doing *A* essentially *involves* an expectation or demand that *S* not have done *A*.

paper, Widerker (2005) suggests that the appeal of his principle stems from the commonsense idea that "when we consider someone morally blameworthy for a certain act, we do so because we believe that morally speaking he should *not* have done what he did" (p. 296). However, there are reasons to doubt this explanation of why his principle (or my revised version of it) is attractive. This is because on the most natural reading of what it means to believe that someone should not have acted as he did, the fact that we blame a person because we believe this about him does *not* in fact support his principle (or my revised version) at all.

To illustrate, consider a case in which an individual, Jones, lies for some personal gain. Suppose we believe that Jones should not have lied. What is it that we believe? On the most natural construal, to believe that Jones should not have lied is to believe that, morally speaking, he *ought not* to have lied – that is, that he had a moral obligation not to lie which he violated through his action. However, if this is what is meant by our belief that Jones should not have lied, then it is far from clear that this belief supports the claim that it would be reasonable to expect or demand that Jones not have lied. This is because not only is the fact that, morally speaking, a person ought not to have done something *distinct* from the fact that it would be reasonable to expect or demand that he not have done it; it is also the case that believing that, morally, he ought not to have done it does not *entail* that it would be reasonable to expect or demand that he not have behaved that way.

Curiously, Widerker (2005) himself recognizes this difference and lack of entailment. However, he does *not* apparently appreciate the effect this has on undermining his claim about what explain his principle's appeal. In his paper, Widerker

provides an example in which though it is the case that, morally speaking, a person ought not to have done what he did, the particular circumstances entail that it is *not* reasonable to expect or demand of him that he not have acted that way. He describes an individual who promised to return a book to a friend by a certain time but, because of circumstances beyond his control, is unable to do so (p. 303). Widerker argues that the fact that he cannot return the book does nothing to remove the person from his moral obligation to return it by the particular time. After all, the person might say, "I know I ought to return it, but I can't do so." However, this fact *does* ensure that it would not be reasonable to expect or demand that he not have acted as he did and instead give back the book because this was not something that it was within his power to do.

I suspect that part of Widerker's error here is to think that he needs to explain the attractiveness of his principle by showing how it is apparently entailed by facts about why we consider people to be blameworthy for what they do. But I do not think that there is any need for him to do this in order to show that his principle is appealing. Instead, why can't we just say that his principle (or my revised version of it) is attractive because it offers a natural and compelling way of capturing the intuitive connection that exists between moral blame on the one hand and moral expectations or demands on the other? It is unclear to me why the appeal of his principle would need any more explanation than this.[37]

---

37 As I suggested earlier, Widerker can argue that his principle is more attractive than a similar, but stronger, claim about the relationship between blame and expectations or demands – namely, that a person's blameworthiness *involves* an expectation or demand that he not act as he did – on the following grounds: it is more conceivable to blame an individual without demanding or expecting anything of him than it is to conceive of blaming him without it being reasonable *to* form a demand or expectation of him.

7.3 The Authorship assumption.

Having outlined and defended the reasonable demand argument, I now want to make good on an earlier claim I suggested and explain *why* this argument is so significant in recasting the traditional debate about freedom and responsibility. The argument is important in this respect because it draws attention to what I see as a traditional assumption about the relationship between freedom and moral responsibility. Let me explain.

While philosophers have paid a great deal of attention to the issue of what *kind* of freedom people must act with in order to be morally responsible for what they do, far less attention has been paid to the important, but neglected, question of why freedom is even an *issue* when it comes to determining a person's moral responsibility. Why is it apparently the case that people must act with a certain kind of freedom in order to merit praise or blame for their behavior? In my view, the usual, or assumed, answer to this question reflected in most of the work on freedom and responsibility is that the issue of a person's freedom is relevant to determining his moral responsibility simply because people must act with a certain kind of freedom in order for them to bring about or 'author' their actions, moral responsibility requiring this kind of authorship. On this view, once we figure out the kind of freedom with which a person must act in order for his behavior to be attributable to him, there are no more freedom-relevant questions that need to be asked, no more reasons to be interested in his freedom, in order to determine his moral responsibility. Call this:

The Authorship Assumption:

> The issue of a person' freedom as it pertains to his moral responsibility is simply the issue of determining the kind of freedom needed for him to bring about or author what he does.[38]

Reflecting a commitment to this assumption, the traditional battle between those who accept PAP and those who reject it has centered upon whether or not people must have alternative possibilities in order for their actions to be attributable to them. On the one hand, PAP-adherents have generally argued that the reason that moral responsibility requires the freedom to do otherwise is because unless people meet a genuine fork in the road when acting, their behavior cannot *truly* be said to be something they brought about or authored. On this view, it is only when people have the freedom to do otherwise that they can leave their distinctive mark or 'stamp' on the world, and so only then that their behavior can be fully *theirs*.

To give just one instance of this justification for PAP, consider Keith Wyma's (1997) example from his childhood of when he learned to ride a bike. When he first rode the bike without assistance, it was, he writes, "because I had the leeway to falter but did not do so, [that] the success of riding was truly mine" (p. 68). According to Wyma, what made it the case that the bike-riding was something *he* did, an event that could properly attributed to him, was that he had the freedom to do otherwise – the freedom to fall over, for instance. Without this sort of freedom, Wyma (and the majority of other PAP-

---

38 As I mentioned a few sentences ago, the authorship assumption rests on a background claim – which I will accept, for the sake of argument – that a person can be morally responsible only for actions that are attributable to him, only for actions of which he is the author.

defenders) argue that his bike-riding success would not be something he really did, and so could not be something for which he could be praised.

However, on the other hand, a natural way to look at the new compatibilist conditions of Frankfurt, Fischer, Watson, and others, is as different ways of developing the claim that people *can* author their behavior without possessing the freedom to do otherwise. I think that it is plausible to argue that the reason these compatibilists draw a difference between those occasions when a person's actions reflect his 'real' self and those times when they do not, and between those times when his behavior is responsive to reasons and those occasions when it is not, is as a way to capture the sense in which people's actions can sometimes be attributed to them as things *they* did, things they fully brought about.

Derk Pereboom (2001) gestures at a similar interpretation of these compatibilist conditions. He describes them as attempts to "tie moral responsibility to actions that are in some way or the other causally integrated with features of an agent's psychology" (p. 100). We can push Pereboom's interpretation further and ask *why* compatibilists would be interested in developing this kind of integration at all. In my view, the natural answer is that they want to capture the sense in which people can bring about or author their actions without possessing the freedom to do otherwise.[39]

---

39 Appealing to the authorship assumption helps to make sense of some remarks in a recent paper by Fischer (2006). Calling the freedom to do otherwise 'regulative control,' and the other kinds of freedom not reliant on alternative possibilities 'guidance control,' Fischer seeks to "provide a measure of intuitive support for the claim that guidance control is all the control (or freedom) necessary for moral responsibility" (p. 107) by attempting to "identify … the 'picture' that supports the claim that guidance control, and not regulative control, is required for moral responsibility" (p. 117). He argues that those who argue that PAP is true are motivated by, what he calls, the 'making a difference'

However, the fact that both those who accept PAP and those who reject it have tended to accept the authorship assumption has led to a stalemate with respect to establishing PAP's truth or falsity. For it is unclear how to settle the question of whether the freedom to do otherwise is required for people to author their behavior without falling into a battle of competing intuitions. That is, naturally-inclined PAP-rejecters will find the new compatibilist conditions an attractive way of tying an individual to his actions, while PAP-sympathizers will insist that without meeting a genuine fork in the road, people cannot *really* author what they do.

It is in this context that the reasonable demand argument is so important. It, and the principle on which it rests, PAD (the principle that a person is blameworthy only if it would be reasonable to demand that he not have acted as he did), puts PAP-defenders in a position of strength. It offers them a way to sidestep the controversy about whether the

---

picture. That is, the intuition that drives their accounts of moral responsibility is that "being morally responsible involves making a certain sort of difference to the world. If you make a difference, in this sense, you *select* which path the world will take, among various paths that are genuinely available. Your selection determines which way the world goes, and you thereby make a crucial difference" (p. 121). However, the 'making a difference' intuition of moral responsibility is not available to new compatibilists for they deny that responsibility requires the freedom to do otherwise. Instead, Fischer suggests that these compatibilists are motivated by a different intuitive picture of responsibility on which it is 'making a statement,' rather than making a difference, that is crucial for morally responsible action. On this alternative view, responsibility is a matter of making a statement in the sense that a person "engage[s] in a particular kind of self-expression" (p. 117) when he acts. However, I suspect that the reason each party might talk about making a difference or making a statement in the first place is simply as a way to capture the sense in which morally responsible individuals truly author what they do. That is, I think that if PAP-sympathizers do think of themselves as motivated by the intuition that responsible agents 'make a difference' to the world, they do so because they think that difference-making is required for authorship of one's actions. On the other hand, if Fischer is right to think that new compatibilists are motivated to describe moral responsibility in terms of 'making a statement,' or self-expression, they do so because

freedom to do otherwise is required for authorship of one's behavior. In fact, armed with the reasonable demand argument, PAP-adherents can grant, for the sake of argument, that the new compatibilist conditions *do* capture the freedom required for people's actions to be attributed to them without thereby conceding that blameworthiness does not require alternative possibilities. For PAD shows that the authorship assumption is false, and there is *more* to delineating the freedom required for moral blame than simply determining the freedom needed for people to bring about their actions. In particular, in order for a person to be blameworthy, we must *also* ask whether he acted with sufficient freedom for it to be reasonable to demand that he not have acted as he did. As we saw earlier, in order for this kind of demand to be fair or reasonable, it must the case that the individual could have done otherwise.

In other words, the reasonable demand argument's significance rests on the fact that even if we were to grant new compatibilists the claim that people can author their behavior without the freedom to do otherwise, this would *not* show that alternative possibilities are not needed for blame. For, as the argument shows, there is a *further* reason to think about the freedom with which a person acts when determining his blameworthiness than simply whether he acts with enough freedom to author his behavior. Moreover, this additional freedom consideration – of whether the individual acts with sufficient freedom for it to be reasonable to demand that he not have acted that way – supports the claim that moral blame really *does* require alternative possibilities after all, whether or not it is also needed for people to author what they do. With its

---

they think that statement-making is all that is needed for people's actions to be truly attributed to them as things they bring about.

rejection of the authorship assumption, the dialectical move involved in the reasonable demand argument represents a significant shift not only in defenses of PAP but also in our more general understanding of the relationship between freedom and moral responsibility.

References

Arpaly, N. & Schroeder, T. (1999). Praise, blame, and the whole self. *Philosophical Studies*, **93**, 161-188.

Austin, J. L. (1979). *Philosophical Papers*. Oxford: Oxford University Press.

Ayer, A. J. (1954). *Philosophical Essays*. London: Macmillan.

Baker, Rudder, L. (2006). Moral responsibility without libertarianism. *Nous*, **40**, 307-330.

Capes, J. (2009). The W-defense. Forthcoming in *Philosophical Studies.*

Chisholm, R. (1964). Human freedom and the self. *The Lindley Lectures*, University of Kansas.

Ekstrom. L. W. (2000). *Free Will*. Boulder, CO: Westview.

Ekstrom, L. W. (2005). Alienation, autonomy, and the self. *Midwest Studies in Philosophy*, **29**, 45-67.

Fischer, J. M. (1986). [Ed.] *Moral Responsibility*. Ithaca: Cornell University Press.

Fischer, J. M. (1999). Recent work on moral responsibility. *Ethics*, **110**, 93-139.

Fischer, J. M. (2006). *My Way*. Oxford: Oxford University Press.

Fischer, J. M., Kane, R., Pereboom, D. & Vargas, M. (2007). *Four Views on Free Will*. Oxford: Blackwell.

Fischer, J. M. & Ravizza, M. (1998). *Responsibility and Control*. Cambridge: Cambridge University Press.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, **66**. 829-839.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, **68**, 5-20.

Frankfurt, H. (1975). Three concepts of free action. *Proceedings of the Aristotelian Society Supplementary Volume*, **49**, 113-125.

Frankfurt, H. (1976). Identification and externality. In Rorty, A. (Ed.) *The Identities of Persons*. Berkeley: University of California Press.

Frankfurt, H. (1992). The faintest passion. *Proceedings and Addresses of the American Philosophical Association*, **66**, 5-16.

Frankfurt, H. (1999). *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.

Frankfurt, H. (2002). Response to Bratman. In Buss, S. & Overton, L. (Eds.) *Contours of Agency*. Boston: MIT Press.

Frankfurt, H. (2003). Some thoughts concerning PAP. In Widerker, D. & McKenna, M. (Eds.) *Moral Responsibility and Alternative Possibilities*. Aldershot, UK: Ashgate.

Ginet, C. (1996). In defense of the principle of alternative possibilities: Why I don't find Frankfurt's argument convincing. *Philosophical Perspectives*, **10**, 403-417.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.

Ginet, C. (2002). Review of Pereboom's *Living Without Free Will*. *Journal of Ethics*, **6**, 305-309.

Ginet, C. (2003). Reprint of Ginet 1996 with Addendum discussing Mele & Robb 1998. In Widerker, D. & McKenna, M. (Eds.) *Moral Responsibility and Alternative Possibilities*. Aldershot, UK: Ashgate.

Ginet, C. & Palmer, D. On Mele and Robb's indeterministic Frankfurt-style case. Forthcoming in *Philosophy and Phenomenological Research*.

Haji, I. & McKenna, M. (2004). Dialectical delicacies in the debate about freedom and alternative possibilities. *Journal of Philosophy*, **101**, 299-314.

Haji, I. & McKenna, M. (2006). Defending Frankfurt's argument in deterministic contexts: A reply to Palmer. *Journal of Philosophy*, **103**, 363-372.

Kane, R. (1985). *Free Will and Values*. Albany: State University of New York Press.

Kane, R. (1996). *The Significance of Free Will*. Oxford: Oxford University Press.

Kane, R. (2005). *A Contemporary Introduction to Free Will*. Oxford: Oxford University Press.

King, M. (2009). Review of Haji's *Incompatibilism's Allure*. *Notre Dame Philosophical Reviews*. URL = < http://ndpr.nd.edu/review.cfm?id=15546>

Lippert-Rasmussen, K. (2003). Identification and responsibility. *Ethical Theory and Moral Practice*, **6**, 349-376.

Locke, D. (1975). Three concepts of free action. *Proceedings of the Aristotelian Society Supplementary Volume*, **49**, 95-112.

McKenna, M. (2004). Compatibilism. *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition) Zalta, E. N. (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/compatibilism/>.

McKenna, M. (2005a). Where Frankfurt and Strawson meet. *Midwest Studies in Philosophy*, **29**, 163-180.

McKenna, M. (2005b). The relationship between autonomous and morally responsible

agency. In J. S. Taylor (ed.) *Personal Autonomy*. Cambridge: Cambridge University Press.

McKenna, M. (2008a). Frankfurt's argument against alternative possibilities: Looking beyond the examples. *Nous*, **42**, 770-793.

McKenna, M. (2008b). A hard-line reply to Pereboom's four-case manipulation argument. *Philosophy and Phenomenological Research*, **77**, 142-159.

Mason, E. (2005). Recent work on moral responsibility. *Philosophical Books*, **46**, 343-353.

Mele, A. & Robb, D. (1998). Rescuing Frankfurt-style cases. *Philosophical Review*, **107**, 97-112.

Mele, A. & Robb, D. (2003). Bbs, magnets, and seesaws: The metaphysics of Frankfurt-style cases. In Widerker, D. & McKenna, M. (Eds.) *Moral Responsibility and Alternative Possibilities*. Aldershot, UK: Ashgate.

Moore, G. E. (1912). *Ethics*. London: Williams and Norgate.

Palmer, D. (2005). New distinctions, same troubles: A reply to Haji and McKenna. *Journal of Philosophy*, **102**, 474-482.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

Pererboom, D. (2005). Defending hard incompatibilism. *Midwest Studies in Philosophy*, **29**, 228-247.

Pereboom, D. (2008). A hard-line reply to the multiple-case manipulation argument. *Philosophy and Phenomenological Research*, **77**, 160-170.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge: Harvard University Press.

Smith, A. (2005). Responsibility for attitudes. *Ethics*, **115**, 236-271.

Smith, A. (2008). Control, responsibility, and moral assessment. *Philosophical Studies*, **138**, 367-392.

Stump, E. (1988). Sanctification, hardening of the heart, and Frankfurt's concept of the will. *Journal of Philosophy*, **85**, 395-420.

Van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

Watson, G. (1975). Free agency. *Journal of Philosophy*, **72**, 205-220.

Watson, G. (1987). Free action and free will. *Mind*, **96**, 145-172.

Watson, G. (1999). Soft libertarianism and hard compatibilism. *Journal of Ethics*, **3**, 351-365.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, **24**, 227-248.

Watson, G. (2004). *Agency and Answerability*. Oxford: Oxford University Press.

Widerker, D. (1995). Libertarianism and Frankfurt's attack on the principle of alternative possibilities. *Philosophical Review*, **104**, 247-261.

Widerker, D. (2000). Frankfurt's attack on the principle of alternative possibilities: A further look. *Philosophical Perspectives*, **14**, 181-201.

Widerker, D. (2005). Blameworthiness, non-robust alternatives, and the principle of alternative expectations. *Midwest Studies in Philosophy*, **29**, 292-306.

Widerker, D. (2006). Libertarianism and the philosophical significance of Frankfurt scenarios. *Journal of Philosophy*, **103**, 163-187.

Widerker, D. & McKenna, M. [Eds.] (2003). *Moral Responsibility and Alternative Possibilities*. Aldershot, UK: Ashgate.

Wolf, S. (1987). Sanity and the metaphysics of responsibility. In Schoeman, F. [Ed.] *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press.

Wolf, S. (1990). *Freedom Within Reason*. Oxford: Oxford University Press.

Wyma, K. (1997). Moral responsibility and leeway for action. *American Philosophical Quarterly*, **34**, 57-70.

Zimmerman, D. (1981). Hierarchical motivation and freedom of the will. *Pacific Philosophical Quarterly*, **62**, 354-368.

Vita

David William Palmer grew up in Worcester, England. He received the degree of Bachelor of Science from the University of York in 2001. He then studied at King's College London, gaining the degree of Master of Arts in 2002. In 2003, he entered the Graduate School at the University of Texas at Austin. In 2009, he took up a position as Lecturer in Philosophy at the University of Tennessee, Knoxville.

Permanent Address: 89 Malvern Road, Worcester, WR2 4LJ, UK.

The Dissertation was typed by the author.